COMPUTATIONAL APPROACHES FOR ASSESSING KINOME FUNCTION AND DEREGULATION

A Dissertation

Presented to the Faculty of the Weill Cornell Graduate School

of Medical Sciences

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Charles Joseph Murphy

August 2018

COMPUTATIONAL APPROACHES FOR ASSESSING KINOME FUNCTION AND DEREGULATION

Charles Joseph Murphy, PhD

Cornell University 2018

Protein kinases are a diverse family of about 500 proteins that all share the common ability to catalyze phosphorylation of the side chains of amino acids in proteins. Kinases play a vital role across diverse biological functions including proliferation, differentiation, cell migration, and cell-cycle control. Moreover, they are frequently altered across most cancers types and have been a focus for development of anti-cancer drugs, which has led to the development of 38 approved kinase inhibitors as of 2018. In this thesis, I developed two orthogonal computational approaches for investigating kinase function and deregulation. Starting with data from a large cohort of mouse triple negative breast cancer (TNBC) tumors, I use a combination of whole exome sequencing (WES) and RNA-seq to identify somatic alterations that drive individual tumors. I discovered that a large number of these alterations involve protein kinases and subsequent therapeutic targeting led to tumor regression. For my second approach, I utilized a large peptide library dataset from about 300 kinases. Which kinase phosphorylate which phosphorylation site is determined by both kinase-intrinsic and contextual factors. Peptide library approaches provide kinase-intrinsic amino acid specificity, which I used to predict novel kinase substrates and map out kinase phosphorylation networks. In summary, I developed methods using next-

generation sequencing and peptide library data to generate novel insights into protein kinase function and deregulation.

BIOGRAPHICAL SKETCH

Charlie graduated from University of Wisconsin-Milwaukee in 2013 with a Bachelor of Science in Mathematics. Throughout his undergraduate career he participated in various research experiences, but the pivotal one was his acceptance into the two-year fellowship, Undergraduate Research Experiences in Aquatic Biology and Mathematical Modeling, that inspired his interest in computational biology. He enrolled in the Tri-Institutional Training Program in Computational Biology and Medicine at Weill Cornell Medical College in 2013 and joined the labs of Dr. Lewis Cantley and Dr. Olivier Elemento.

ACKOWLEDGEMENTS

I would like to thank my thesis advisors Lewis Cantley and Olivier Elemento for their guidance and patience. I would also like to thank lab members of both the Cantley and Elemento labs for the advice and training they offered on a variety of questions and problems I posed. I especially want to thank Hui Liu and Jared Johnson for their collaboration that enabled me to achieve my dissertation. I would also like to thank the Tri-I CBM program: David Christini for his advice and Margie Hinonangan-Mendoza for the reminders and gentle pushes to meet my deadlines.

I am also grateful for my various undergraduate advisors; without the opportunity they provided I would never have pursued a PhD. Gabriella Pinter and Istvan Lauko, who accepted and guided me in their undergraduate fellowship that sparked my interest in computational biology. Peter Tonellato, who let me work in his UWM and Harvard labs my last couple undergraduate years. He not only taught me science but also inspired me to have more confidence and persistence.

Finally, I would like to thank my partner, Mike, and my family for their consistent support.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Kinases are crucial components for the normal functioning of cell biology (Manning *et al.*, 2002; Weinberg, 2013). Kinases are a diverse set of proteins which all have a conserved catalytic domain of ~250 amino acids that catalyzes the phosphorylation reaction, the transfer of a phosphate from ATP to a serine (S), threonine (T), or a tyrosine (Y). There are at least 518 known kinases in the human genome and hundreds of thousands of known phosphorylation sites. The phosphorylation event is a vital signaling processing message that all cells use. Protein phosphorylation can enhance or impair protein interactions, activate or inactivate protein enzymatic activity, enhance or impair protein degradation or facilitate or impair re-localization to a different cellular compartment. Many of the known kinases have therefore been discovered to play vital roles in proliferation, differentiation, cell migration, and cell-cycle control.

Kinases are also frequent causes of tissue dysfunction, notably cancer (Manning et al., 2002). In fact, the first gene to be discovered that could drive cancer was the tyrosine kinase, SRC (Weinberg, 2013). Since then, dozens of kinases have been implicated in driving cancer, leading to the development of 38 approved kinase inhibitors as of 2018 (Ferguson and Gray, 2018; Lemmon and Schlessinger, 2010). There are numerous mechanisms by which kinase function can be derailed. Mutant kinases can be deactivated or be made constitutively activated. The expression level of the kinase can be altered via genomic amplification or deletion as well as via

deregulation of gene regulatory networks. Gene fusions can pair the kinase with new protein function or regulation elements that affects the cellular localization or activity level of the kinase. Detecting and studying this plethora of ways by which kinases can be altered requires numerous types of technology.

1.1 Kinase signaling networks underpin normal cell functioning

The eukaryotic protein kinases domain contains two parts: an N-terminal lobe of β-sheets and C-terminal lobe of α-helices (Ubersax and Ferrell, 2007). Despite the highly conserved catalytic domain, there are functionally significant ways to group the kinases. One of the most important ways is based on whether they phosphorylate tyrosine or serine/threonine. Histidine, lysine and aspartate phosphorylation also occur in eukaryotic cells, though in most cases these modifications are intermediates in enzymatic reactions rather than transphosphorylation by a protein kinase. Only 20% of kinases are tyrosine kinases, differentiated from serine/threonine kinases by a set of highly conserved residues that provide room for the large, aromatic phosphoacceptor (tyrosine) (Taylor *et al.*, 1995; Chen *et al.*, 2014). Serine/threonine kinases contain conserved residues that allow for a small, aliphatic phosphoacceptor (serine or threonine). Moreover, some serine/threonine kinases show preference for either serine or threonine based on the presence of a single residue in the kinase activation segment termed "DFG+1" residue (Chen et al., 2014). Furthermore, the proportion of known phospho-tyrosines (pY), phospho-serines (pS), and phospho-threonines (pT) that have been identified reflects the relative number of protein kinases with selectivity for each

of the three amino acids. Generally, pS makes up ~86%, pT makes up ~12%, and pY makes up ~2% of known phosphorylation sites (Newman *et al.*, 2014). The somewhat disproportional abundance of pS can be accounted for by the fact that some Ser kinases have dozens or even hundreds of substrates.

Kinases can be further grouped into families based on evolutionary and functional relationships. A phylogenetic tree constructed based on the kinase domain demonstrates that kinases can be broken down into families, such as the SRC, FGFR, and Eph families (Figure 1-1). The FGFR family for example consists of four members: FGFR1, FGFR2, FGFR3, and FGFR4. These four tyrosine kinases are structurally similar and are critical in embryogenesis and organogenesis, as well as mediating metabolic functions, tissue repair, and regeneration in adult tissues (Ornitz *et al.*). Although kinases from the same family often share similar function, outside the conserved kinase domain are a diverse set of other regulatory and catalytic domains that further specify kinase function. These enable kinase-specific tissue- and cellular-specific expression patterns and placement within protein interaction networks. One example is the distinction between receptor and non-receptor tyrosine kinases. Receptor tyrosine kinase (RTKs) all contain an extracellular ligand domain, a single transmembrane helix, and a cytoplasmic region containing the kinase domain and other regulatory domains (Lemmon and Schlessinger, 2010). There are 58 human RTKs and these are an important class of signaling kinases for cellular proliferation, differentiation, metabolism, and more. Importantly, RTKs are frequently involved in a variety of human diseases.

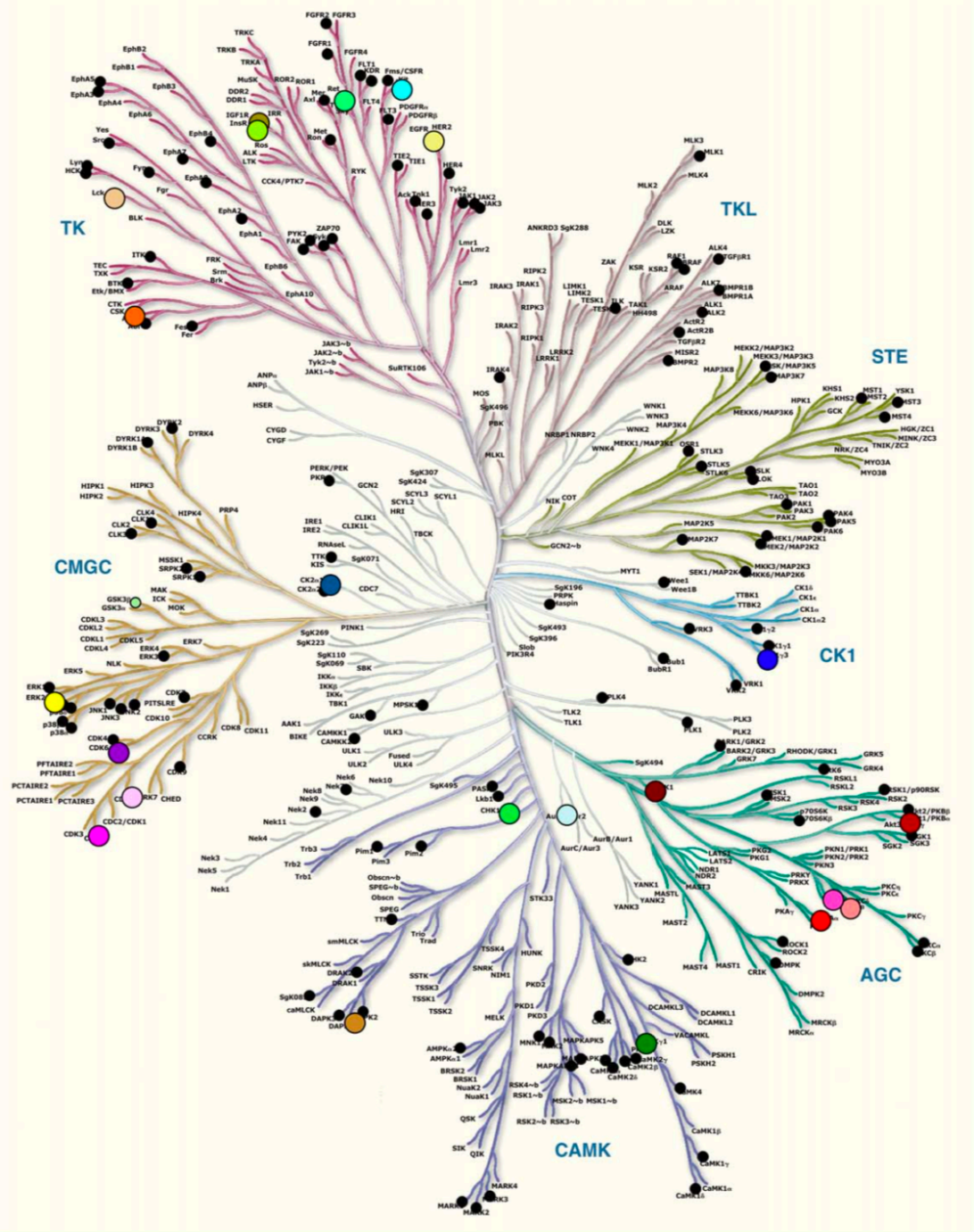

Figure 1-1: **Family tree of the human kinome.**
Reprinted with permission from Trends in Biochemical Sciences (Taylor and Kornev, 2011). Shows the family groupings of the protein kinases.

Phosphorylation is responsible for most of cell signal transduction, which effects such fundamental processes as cell proliferation, movement, apoptosis, and differentiation. With over 500 kinases and hundreds of thousands of known phosphorylation sites, which kinase phosphorylates which serine, threonine, and tyrosine is determined by both kinase-intrinsic and contextual factors (Figure 1-2) (Hornbeck *et al.*, 2004; Dinkel *et al.*, 2011). Although all kinases have the conserved kinase catalytic cleft, the surface charge and hydrophobicity around the binding pocket where the substrate binds can vary. This leads to each kinase preferring certain types of amino acids C-terminal and N-terminal from the central serine, threonine, or tyrosine (Ubersax and Ferrell, 2007). Additional contextual factors such as cellular localization, protein-protein interaction networks, tissue expression, intrinsic instability of the substrate protein, and others factors determine whether a site can be phosphorylated and by which kinase.

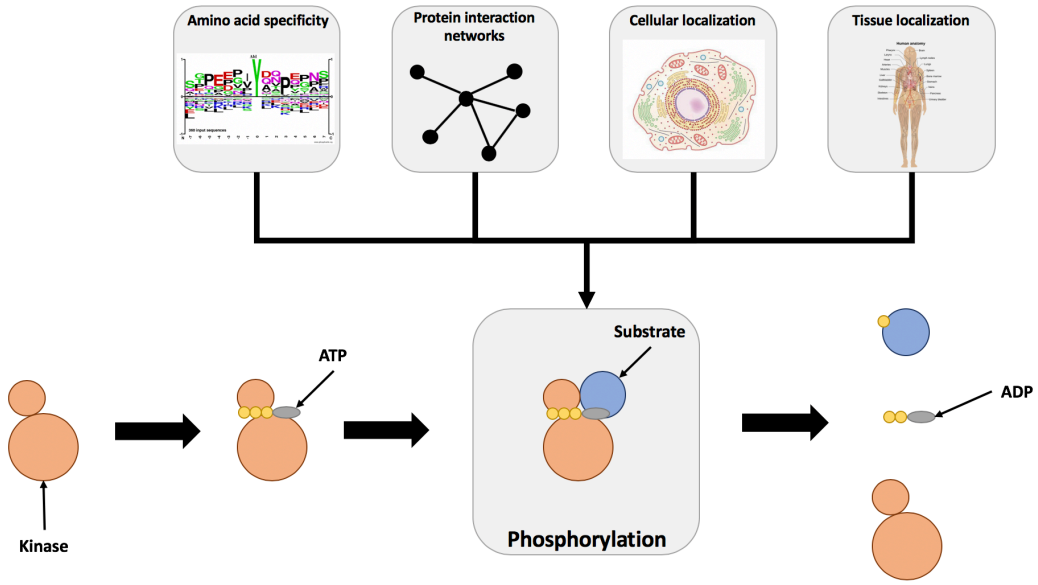

Figure 1-2**: Determinants of kinase phosphorylation.**
The top four squares show different contextual factors that influence which substrates a kinase can phosphorylate. The bottom of the figure outlines the phosphorylation reaction.

1.2 Kinases are frequent drivers of cancer

Recent large-scale cancer studies have revealed the extent of involvement of kinases in human cancer. Lawrence *et al*. identified 260 recurrently point mutated genes across 21 tumor types, and 27 (10.3%) of them are in protein kinases (Figure 1-3) (Lawrence et al., 2014). This is significant considering that only ~2% of human genes are protein kinases. Yoshihara *et al*. examined gene fusions across 4366 tumors from 13 tumor types and found that 324 (7.4%) of samples contained in-frame fusions involving protein kinases (Yoshihara et al., 2015). Furthermore, Stransky *et al*. examined fusion kinases that occur in ≥ 2 samples across 6,893 tumors and 20 solid tumor types (Stransky *et al.*, 2014). They found that about 3% of samples contain at

least one of the recurrent fusions and conclude there is significant therapeutic potential for targeting them. Beroukhim *et al*. analyzed recurrent somatic copy number alterations across 3,131 samples from 31 histological cancer types and found an enrichment of kinases in amplified regions (Beroukhim et al., 2010). These are other cancer-specific studies repeatedly demonstrate the pervasive involvement of protein kinases in driving cancer.

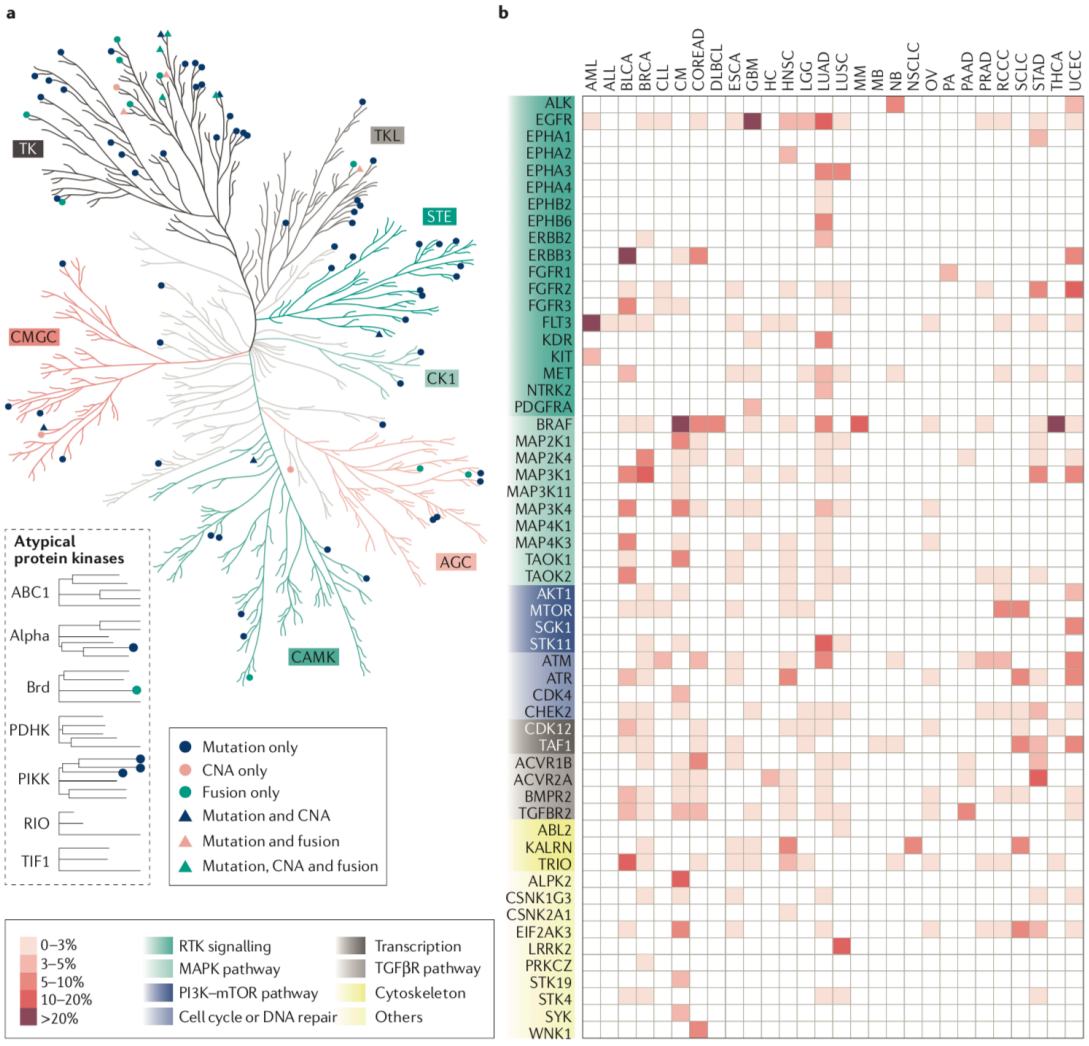

Figure 1-3: **Driver kinases identified from cancer genomic studies.**
Reprinted with permission from Nature Reviews (Fleuren *et al.*, 2016). (A) Shows the kinase family tree annotated with somatic alterations common to particular kinases. (B) Shows the mutation frequency of different kinases by cancer type.

Somatic alterations frequently act indirectly on kinases by being located elsewhere within the kinase signaling pathway (Campbell *et al.*, 2016; Brognard and Hunter, 2011). A recent review not only confirmed that kinases are enriched among the compiled list of 1,100 cancer driver genes, but that 12% of them were substrates of kinases (Fleuren et al., 2016). The PI3K-mTOR and MAPK pathways are some of the most frequently altered pathways in cancer. The main effector protein kinases of the PI3K pathway are the AKT1/2/3 family. Numerous somatic mutations in genes upstream of AKT have been documented including loss of function PTEN deletions, gain of function mutations in PI3K (a lipid kinase), and gain of function mutations or amplifications in AKT1, AKT2 or AKT3. Furthermore, as discussed in Chapter 2, a survey of the genomic landscape of a mouse model of TNBC revealed potential oncogenic alterations in ~50% of the tumors, most of which resulted in the ability to activate protein kinases in the MAPK/PI3K pathways, even though not all genetic aberrations directly affected a known protein kinase (H. Liu et al., 2017).

1.3 The mechanisms of somatic alterations for changing kinase activity

The mechanisms by which kinase function can be altered by a somatic alteration are varied. The somatic alterations can either directly affect kinase activity by altering its genomic sequence, expression level, turnover rate, or its protein substrates. The RTK class of kinases is most often the target of somatic alterations (Fleuren *et al.*, 2016; Lemmon and Schlessinger, 2010). The RTKs generally work via

ligand-induced dimerization and subsequent auto-transphosphorylation of the activation loop of the protein kinase domain. There are four general mechanisms by which alterations can lead to aberrant RTK activity. First, mutations that lead to constitutive activation of the kinase such as KIT mutations found in a variety of cancers that relieve the auto-regulatory inhibition of KIT's tyrosine kinase domain. Second, RTK overexpression either through gene regulation or amplification can lead to ligand-independent dimerization and activation of the RTK as was observed for MET amplification (H. Liu et al., 2017). Third, the kinase domain of the RTK can be fused with another gene leading to its overexpression or constitutive activation. FGFR2-CCDC6 and FGFR2-DNM3 fusions found in breast cancer were shown to be constitutively active due to stable formation of homo-multimerization complexes mediated by the CCDC6 and DNM3 C-terminal fusion partners (H. Liu et al., 2017). Forth, RTKs can become overly active via autocrine activation.

Mutation of the phosphorylation site or its flanking regions is another important mechanism. Reimand *et al.* analyzed somatic mutations from 3,185 tumors across 12 cancer types to assess their impact on known phosphorylation sites. They found that 90% of tumors harbored mutations at or near phosphorylation sites, and predicted that 29% of them either abolish the phosphorylation site or modify it in some way that rewires kinase signaling (Reimand *et al.*, 2013). Several tools have been released that examine the impact of mutations for rewiring phosphorylation signaling networks. Pau *et al.* created ReKINect that predicts three classes of mutations that affect phosphorylation networks (Creixell, Schoof, *et al.*, 2015). First, mutations that directly affect the kinase by making it constitutively "on" or "off".

Second, mutations in upstream or downstream components of a kinase's network. Upstream mutations affect the linear motif surrounding a phosphorylation site, which can lead to a new kinase being able to phosphorylate it. Downstream mutations occur on the kinase and affect its determinates of specificity, which are the key amino acids in the kinase domain that determine its motif preference. Third, mutations that can create or destroy phosphorylation sites. Another tool that examines mutational effects on phosphorylation sites is MIMP (Wagih *et al.*, 2015). In contrast to ReKINect, MIMP predicts just whether a given mutation creates or destroys a phosphorylation site. MIMP was tested on 236,367 mutations from 3,185 tumors across 12 tumor types and revealed that 34,996 of them were within 7 residues of known phosphorylation sites and 7,092 were predicted to either create a site, destroy a site, or switch the phosphorylating kinase.

1.4 In this dissertation

In this dissertation I present analyses from two orthogonal technologies for studying kinase function and deregulation. In chapter 2 I describe the use of RNA-seq and WES to identify mutations, gene fusions, and copy number alterations in a mouse model of TNBC. Through collaboration, I was able to then show that most of the oncogenic alterations involved protein kinases and show that drugs that target the activated protein kinase pathway were effective in treating the cancers. In Chapter 3 I present a strategy to predict proteins substrates of protein kinases utilizing data generated from the oriented peptide libraries to determine the peptide substrate

specificity of a wide range of protein kinases. I demonstrate how next-generation sequencing and peptide libraries are two orthogonal technologies that provide depth of understanding into the function and deregulation of protein kinases in human disease.

CHAPTER 2

ONCOGENIC KINASE DRIVERS IN MOUSE TRIPLE NEGATIVE BREAST CANCER

Breast cancer is one of the leading causes of cancer-related deaths for females in the United States. TNBC is a deadly form of breast cancer defined by the lack of expression of estrogen receptor (ER), progesterone receptor (PR), and HER2 (Foulkes *et al.*, 2010). TNBC is a heterogeneous disease with few recurrent alterations except TP53 alterations (~80%), the PI3K pathway (PIK3CA, PTEN, and INPP4B) alterations, and BRCA1 germline alterations (Shah *et al.*, 2012; Davies *et al.*, 2016). Furthermore, transcriptional profiling of TNBC revealed a number of subtypes each with their own treatment distinctions (Masuda *et al.*, 2013; Lehmann *et al.*, 2011). TNBC is therefore a difficult disease to treat since no single therapy can be designed to target a majority of patients. Instead, treatment must be tailored to individual patients according to their tumor makeup (H. Liu et al., 2017).

Kinases are frequent alterations in breast cancer and TNBC. The Cancer Genome Analysis (TCGA) dataset for mutated genes in 510 breast tumors revealed 3 (8.6%) of the 35 significantly mutated genes were protein kinases (Koboldt, Fulton, *et al.*, 2012). They also confirmed the presence of recurrent copy number alterations in the EGFR, ERBB2, STK11, and MAP2K4 protein kinases. Yoshihara *et al.* found that 105 (8.6%) out of 1,228 breast tumors harbored in-frame protein kinase fusions. Although breast cancer is a very heterogenous disease with many somatic alterations,

with non-protein kinase such TP53, PTEN, and PIK3CA being among the most frequent, many of these alterations occur within protein kinase signaling networks.

In order to study aberrant kinases in TNBC, I used data from a genetically engineered mouse (GEM) model in collaboration with Hui Liu, who performed all the mouse work, molecular assays, and treatment. GEM models are useful for modeling human cancers and exploring treatment options (Frese and Tuveson, 2007). GEM models are one of the more sophisticated animal models for human cancer by allowing mice to be genetically engineered with mutations that predispose them to specific types of cancer. This allows the mouse cancer to be studied throughout the course of its evolution as well as to have various experimental treatments applied to it.

Herein, we utilized a TNBC GEM model in order to characterize the genomic and transcriptomic landscape of mouse and apply precision medicine treatment to cure individual mouse tumors. Our model is the Trp53 $^{flox/flox}$ with our without Brca1$^{flox/flox}$, which have been shown to be strongly predisposed to develop breast cancer with basal-like characteristics (X. Liu et al., 2007)(Figure 2-1). To achieve this, we performed next-generation sequencing on 72 unique mouse primary tumors for our study. In total, we collected we 67 RNA-seq samples from primary tumors, three RNA-seq samples from normal mammary, 63 WES samples from primary tumors, 29 WES samples from paired tail tissue, and 3 WES samples from normal liver tissue. We demonstrated that the TNBC GEM model recapitulates many characteristics of human TNBC, including somatic alteration heterogeneity, low mutation rate, higher gene fusion rate, and similar transcriptional profiles. Finally, we successfully used the mouse model as a precision medicine model for human TNBC by identifying tumor

drivers and targetable alterations. Moreover, we found that many of the oncogenic alterations involved protein kinases. Thus, lending credence to using next-generation sequencing to discover somatically altered kinases.

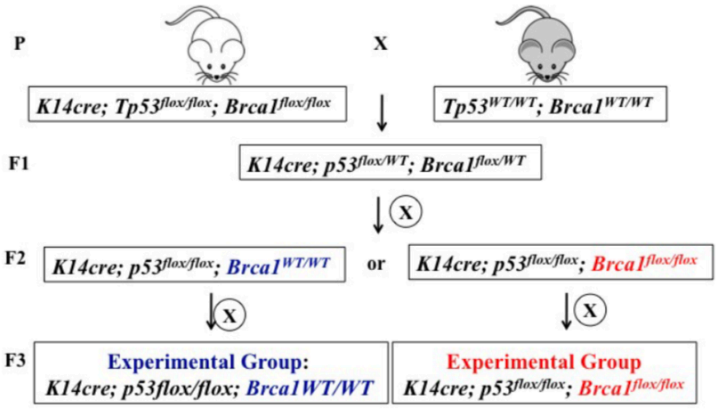

Figure 2-1: **Generation of mouse TNBC tumors.**
Our breeding strategy to generate the K14cre; p53 $^{flox/flox}$; Brca1$^{WT/WT}$ and K14cre; p53 $^{flox/flox}$; Brca1$^{flox/flox}$ mice.

2.1 Modeling human TNBC

Establishing that our mouse model accurately reflects human TNBC is important prior to drawing conclusions that can affect clinical decision making for human patients. Our mouse model is a GEM model, which means the tumors are endogenous to the mouse, in contrast to the commonly used xenograft, which graft human tumors into an immunodeficient mouse (Frese and Tuveson, 2007). There are important limitations to consider for all mouse models, but one important limitation of GEM models is how well they reproduce the kinetics of tumor evolution. Human tumors develop over the course of years or decades, whereas tumors in GEM models usually

develop rapidly with the simultaneous perturbation of one or more driver genes. This can therefore lead one to expect fewer total number of somatic alterations that accumulate. Moreover, mice have longer telomeres compare to humans and will take longer to shorten enough to cause genomic instability, which is an important role in oncogenesis. Hence, assessing the phenotypic and genetic similarity to the human disease counterpart is vital.

Human TNBC is defined as the lack of expression of ER, PR, and HER2 (Foulkes et al., 2010). Thus, we stained a select representative set of the mouse tumors for these receptors (Figure 2-2), demonstrating the tumors are negative for all three. Moreover, previous research has demonstrated that ER, PR, and HER2 status can be determined using RNA-seq. We applied a logistic classifier to each tumor to determine ER, PR, and HER2 status, revealing that the majority of tumors are negative for all three (Figure 2-3).

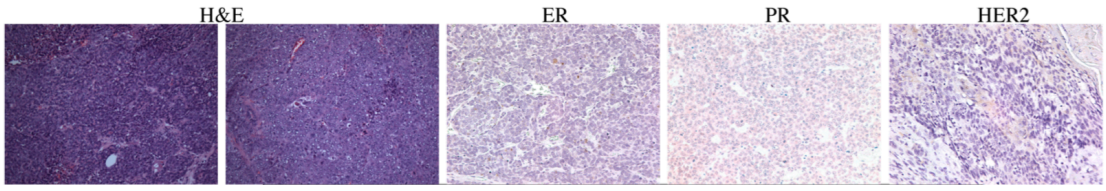

Figure 2-2: **Representative staining for ER, PR, and HER2.**
H&E and IHC staining of ER, PR, and HER2 for tumors.

I next used two published classifiers to determine which breast cancer intrinsic subtypes each tumor falls into. One of the first transcriptional subtype classifiers for breast cancer, PAM50, divided all breast cancers into one of five types: Basal-like,

Luminal A, Luminal B, HER2-enriched, and Normal-like (Parker et al., 2009). These subtypes are associated with various clinical parameters including chemotherapy efficacy and pathological complete response. Prior to classifying the mouse tumors I first converted mouse gene symbols to human gene symbols using the vertebrate homology list provided by Mouse Genome Informatics. I then re-estimated the PAM50 centroids using RNA-seq expression profiles from TCGA BRCA data and PAM50 assignments in the original TCGA BRCA publication as ground truth for cross validation (Koboldt, Fulton, *et al.*, 2012; Parker *et al.*, 2009). After computing the within-sample rank normalization on the RPKM measurements, I achieved 82% classification accuracy after 10-fold cross-validation on the human BRCA TCGA data. I then classified the mouse tumors after rank normalizing the mouse FPKM values. I found that the majority of mouse tumors classify as basal-like (Figure 2-3).

The second classifier I used is the Absolute Intrinsic Molecular Subtype (AIMS) classifier. AIMS uses a large set of rules that examine the relative expression of genes within the sample combined into a Naïve Bayes framework. Prior to classification, I converted mouse gene symbols to human Entrez gene IDs using BioMart and the vertebrate homology list provided by Mouse Genome Informatics (Kinsella *et al.*, 2011; Blake *et al.*, 2017; Paquet and Hallett, 2014). Again, I found that the majority of mouse tumors classified as basal-like (Figure 2-3).

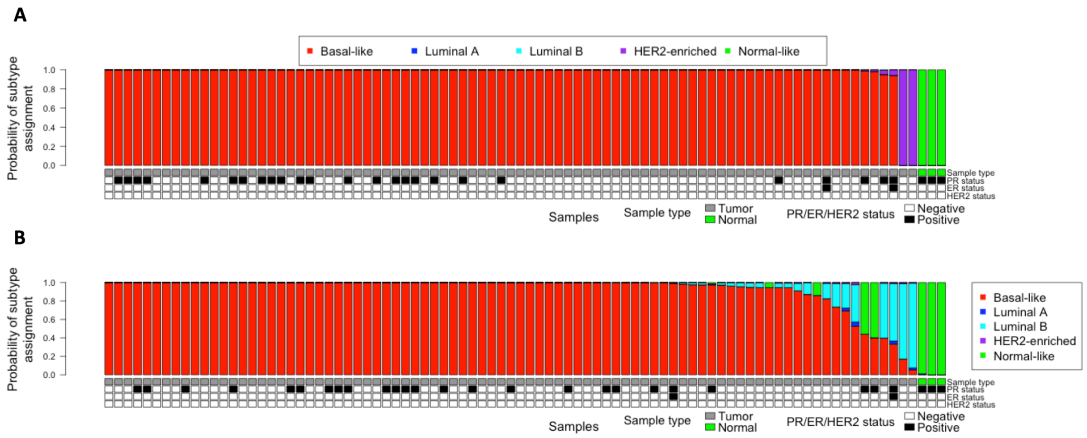

Figure 2-3: **Classification of mouse transcriptomes into TNBC.**
(A) Transcriptional classification of mouse tumors using the AIMS classifier into one of five intrinsic breast cancer subtypes. (B) Transcriptional classification of mouse tumors using the PAM50 classifier into one of five intrinsic breast cancer subtypes.

I further confirmed that our mouse tumors are transcriptionally similar to TNBC by taking an unsupervised approach by clustering (spearman correlation distance, average linkage) on the mouse tumors with breast cancer RNA-seq expression profiles from TCGA on the 137 shared genes between mouse and human that were part of the AIMS signature (Figure 2-4). RNA-seq expression profiles for TCGA breast cancer (BRCA) samples were downloaded from Broad GDAC Firehose (http://gdac.broadinstitute.org/) data version 2016_01_28. I found that the mouse tumors clustered with human TNBC tumors.

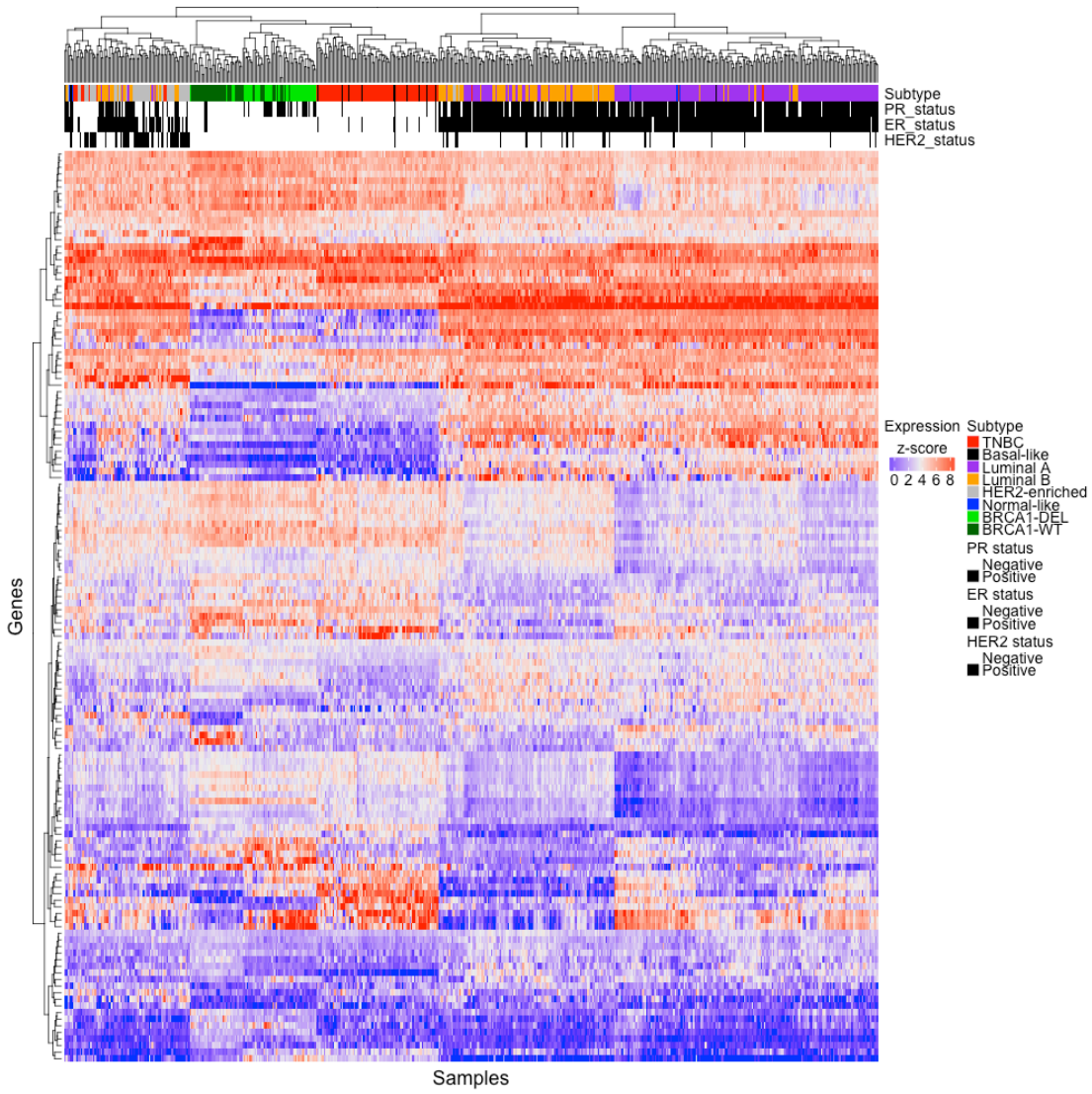

Figure 2-4: **Unsupervised clustering of mouse TNBC with human TNBC.** Mouse TNBC tumors (light and dark green) clustered with human TNBC (red). Genes used for clustering are the genes used in the AIMS classifier.

In summary, I found that our mouse tumors are similar to TNBC. Staining from a representative set of tumors indicate that the tumors are negative for PR, ER, and HER2. Likewise, logistic regression on the RNA-seq indicate most tumors are negative for PR, ER, and HER2. I also classified each mouse tumor according to the intrinsic breast cancer subtypes using two classifiers, finding that the majority of

tumors are basal-like. This finding is significant because 70-80% of human TNBC tumors are basal-like. Finally, I took an unsupervised approach that revealed that our mouse tumors clustered with human TNBC. Hence, I concluded that our mouse tumors model human TNBC.

2.2 Characterization of the mouse TNBC transcriptome and genome

The transcriptome and genome of mouse breast cancer has been previously characterized to a smaller extent using high-throughput technologies. Liu *et al.* first established the K14cre;$p53^{flox/flox}$;$BRCA1^{flox/flox}$ mouse model used in our experiments. Using array CGH for CNAs and microarray for gene expression profiling, they found high rates of genomic instability and similar transcriptional profiles to human breast cancer (X. Liu et al., 2007). Pfefferle *et al.* used a combination of microarray, whole-genome sequencing, and whole-exome sequencing to profile the transcriptome and genome of a number of BALB/c; *Trp53*-null mice (Pfefferle et al., 2016). In addition to finding an unstable genome and being transcriptionally similar to certain human breast cancer subtypes, they identified somatic alterations that were also found in human breast cancer. These include amplifications and deletions such as *Met*, *Cul4a*, *Lamp1*, *Pnpla6*, and *Tubgcp3* that can potentially be targeted by drugs. Ben-David *et al.* used microarray expression profiles to characterize the CNA landscape of a number of breast cancer mouse models, including $Brca1^{-/-}$ and a $p53^{-/-}$ models (Ben-David et al., 2016). They found the various mouse models had significant differences in CNAs profiles, and some recurrent alterations were also found in human cancer. Thus, the

main findings for these studies include genomic instability with high copy number rates, transcriptional profiles similar to human breast tumors, few recurrent alterations, and some overlap with known alterations that occur in human breast cancer with potential as therapeutic targets.

In this thesis, I characterized the most comprehensive set of the mouse breast cancer genomes and transcriptomes to date using combined RNA-seq and whole-exome sequencing on the primary tumors. I identified a heterogeneous number of somatic gene fusions, mutations, and CNAs, many of which involved protein kinases.

2.2.1 Gene Fusions

Gene fusions result from the inter- or intra-chromosomal fusion of DNA strands, that can impair gene function, rearrange gene regulatory elements, or create novel gene products by fusing the coding sequences of two genes. Gene fusions are a major source of driver events across a variety of human cancers, especially breast cancer (Yoshihara *et al.*, 2015; Stephens *et al.*, 2016; Shaver *et al.*, 2016). Among all types of breast cancer, TNBC has one of the highest rates of gene fusion events (Figure 2-5) (Yoshihara et al., 2015). Thus, I sought to identify gene fusions present in our mouse model of TNBC, reasoning they are a likely abundant source of tumor drivers.

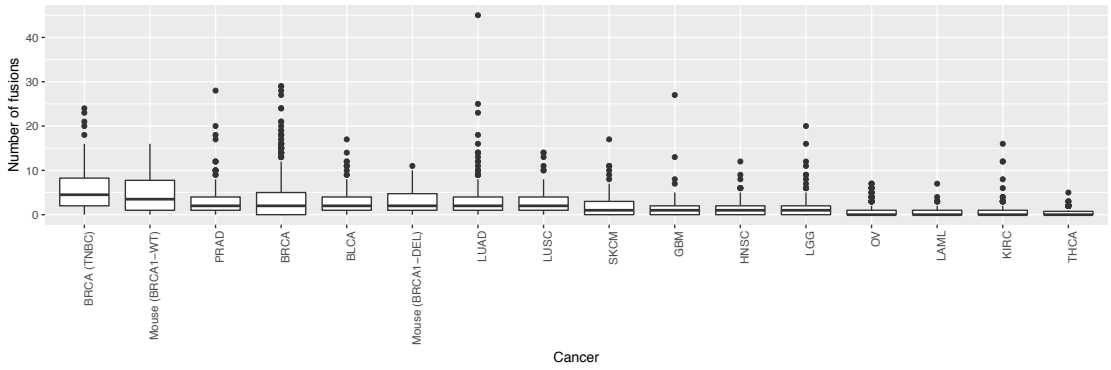

Figure 2-5: **Pan-cancer gene fusion rate.**
Number of gene fusions per tumor across a variety of human cancers (data from TCGA). Human TNBC has the highest rate of gene fusions.

I used FusionCatcher (v0.99.5a) with mouse reference genome GRCm38 to identify gene fusions in our mouse TNBC tumors (Nicorici et al., 2014). I removed likely false positive gene fusions by removing those found in any of the three normal mammary controls, were marked as read-throughs, or occurred with the same breakpoint and gene partners in three or more independent primary tumors. The tumors harbored a very heterogeneous array and number of gene fusions (average of 4 and range of 0 to 16 per tumor), with few being present in more than one tumor (Figure 2-6).

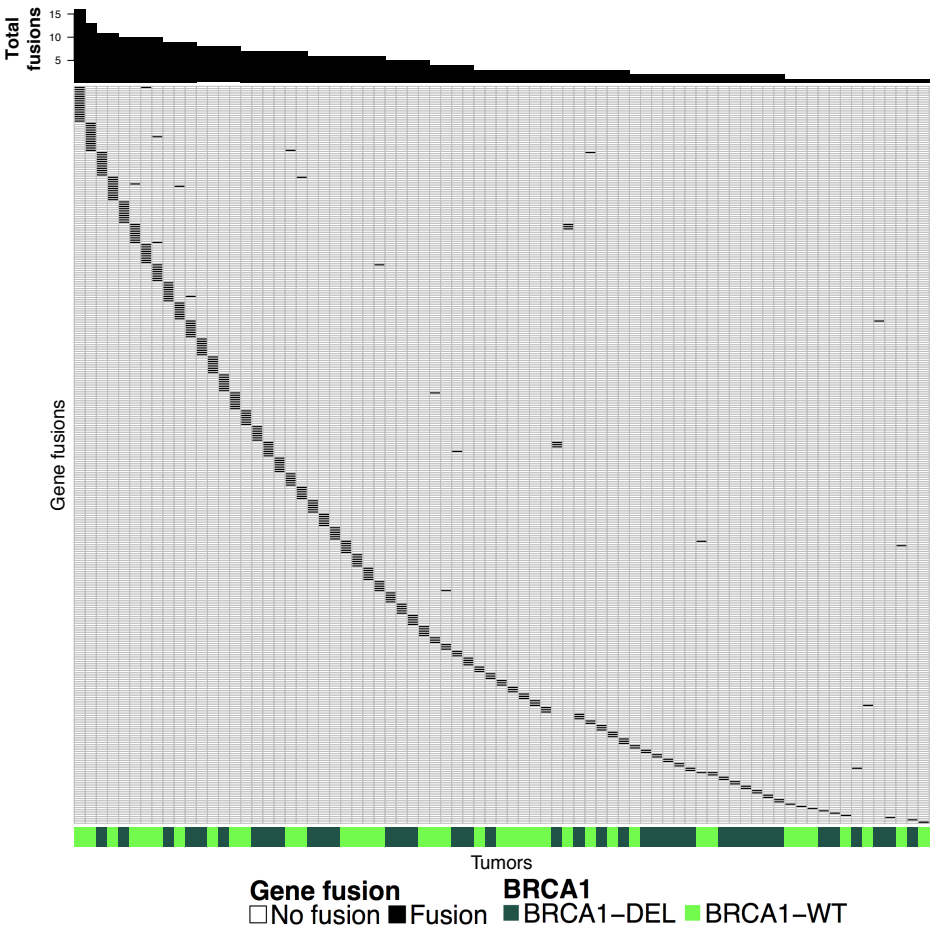

Figure 2-6: **Gene fusions in mouse primary tumors.**
Rows are unique fusions and columns are unique samples. Top graph shows the number of gene fusions per tumor and the bottom shows the BRCA1 genotype.

I found numerous in-frame gene fusions that involved protein kinases with intact kinase domains (Figure 2-7). The kinases involved include *Fgfr2*, *Met*, *Raf1*, and *Braf*. Surprisingly, I found several fusions involving *Fgfr2* as the 5' partner but with different 3' gene partners. In-frame fusions involving *FGFR2*, *BRAF*, *RAF*, and *MET* have all been found previously in human breast cancer, but with different gene partners compared. Several *FGFR2* fusions have been found in breast cancer, which were demonstrated to be sensitive to the FGFR inhibitors PD173074 and pazopanib

(Wu *et al.*, 2013). One fusion involving *BRAF* and *RAF1* each have also been found in breast cancer (Matissek et al., 2018). The *RAF1* fusion was demonstrated to activate downstream signaling and induce cell growth, polarity, and survival on mammary epithelia. The *BRAF* fusion has also been previously found in several lung cancer patients and shown to induce MAPK signaling (Jang et al., 2015). Finally, data from TCGA breast cancer data revealed a patient with a *MET* fusion and another with a *FGFR2* fusion (Yoshihara et al., 2015).

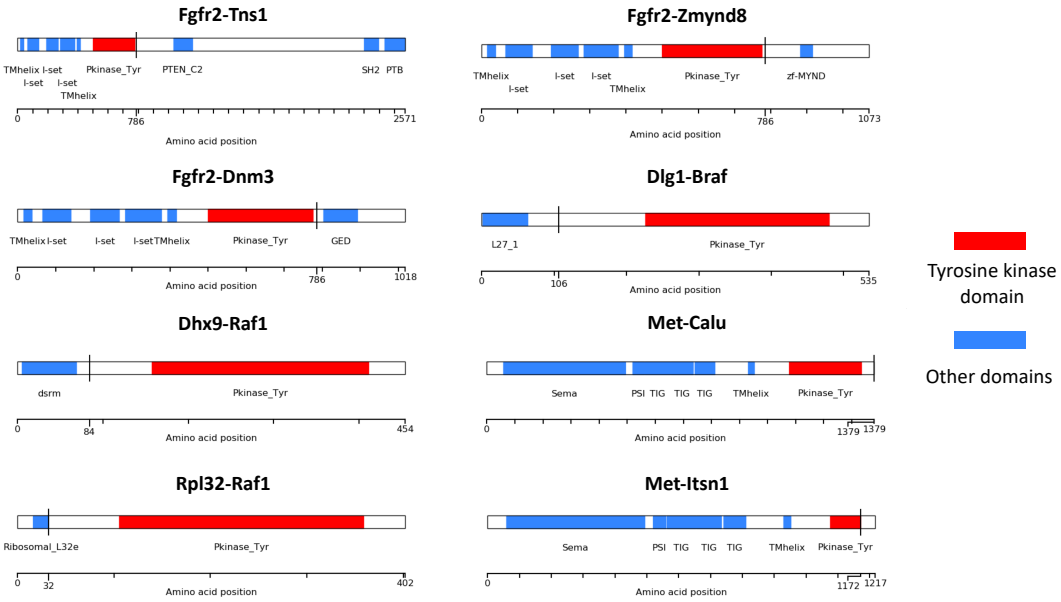

Figure 2-7: **Sample of in-frame kinase fusions.**
Displayed are the protein domain structures of the selected kinase fusions. The gene symbols of the two involved genes are listed at the top. The vertical black line indicates the fusion point. Domain names are listed below the rectangular box.

2.2.2 Copy number alterations

CNAs are regions of a chromosome or whole chromosomes that gain or lose one or more copies, and are very common in cancer (Zack *et al.*, 2013; Beroukhim *et al.*, 2010). Identifying oncogenic CNAs is usually done by looking for genes that are focally and recurrently amplified or deleted and by looking for CNAs that overlap known cancer genes. A recent study examined 4,934 tumor samples from 11 cancer types and found 140 regions that were recurrently and focally subject to CNAs, 102 occur in known cancer genes and 50 in significantly mutated genes. Human breast cancer, especially TNBC, haver higher than average CNA rates and so have very heterogenous CNA landscapes (Figure 2-8).

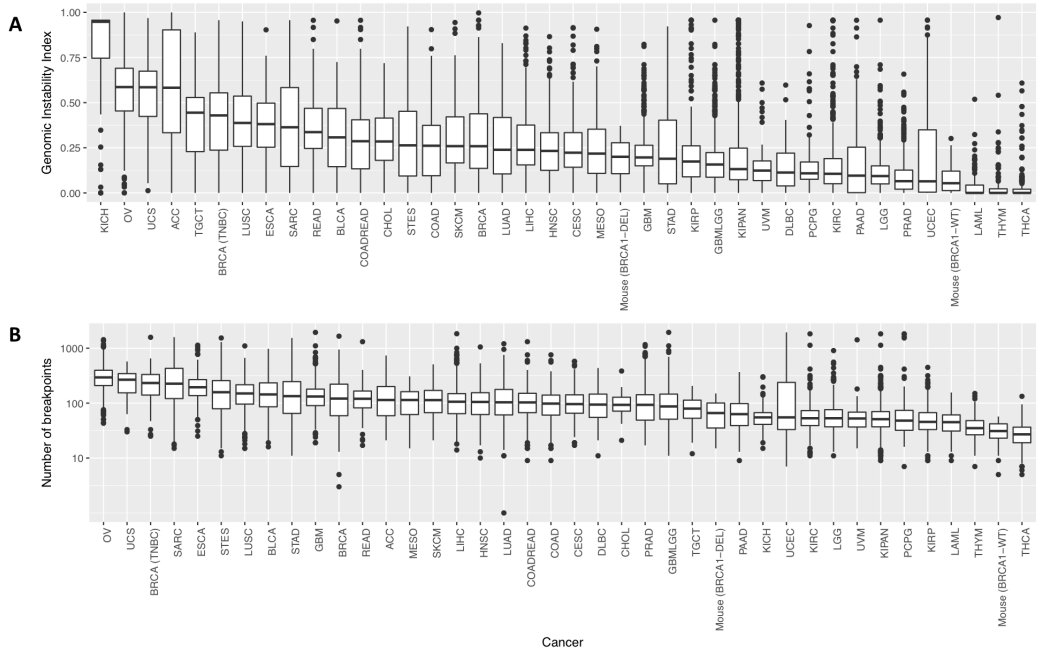

Figure 2-8: **Pan-cancer CNA rates.**
(A) Genomic instability index per tumor across cancer types. (B) Number of breakpoints per tumor across cancer types.

The CNA landscape of our mouse TNBC model using a customized pipeline. Before calling CNAs with CNVkit (v0.8.1) I filtered out reads with mapping quality less than 30 and read pairs that map to different chromosomes using samtools (Talevich *et al.*, 2016). I used separate strategies to call CNAs in tumors sequenced with the SureSelect and NimbleGen kits.

CNAs of tumors sequenced with the SureSelect kit were called using a panel of three livers as controls. However, the built-in GC-content bias correction within CNVkit was not sufficient. I observed increased variance in log2 coverage ratios (tumor over liver) in capture regions with lower GC content (data not shown); therefore, I developed an algorithm to correct the variance. Let L_{ik} be the log2 coverage ratio in capture region k and sample i. Let GC_j denote the set of all capture regions k where GC content is j (in intervals of 1%). Then I compute m_{ij} as the median absolute deviation (MAD) of all L_{ik}, where $k \in GC_j$. I then computed the value M, which is the median of all the m_{ij} values. I then calculated, Lc_{ik}, the corrected log2 coverage ratio with the following equation:

$$Lc_{ik} = \frac{M}{m_{ij}} L_{ik}, where\ k \in GC_j$$

I demonstrated that this correction significantly improves segmentation results as determined by comparison to an independent algorithm, CopywriteR (v2.4), which can call CNAs without the need for a control sample (Kuilman et al., 2015). CopywriteR was not used as the primary CNA analysis because its results are noisier than CNVkit

results but provides an important benchmark by which I demonstrated the validity of our CNA calling strategy. Using the same set of filtered reads as used for CNVkit I ran CopywriteR with window size set to 40kbp and the SD parameter set to 1 for the segmentation step. I compared CNVkit and CopywriteR results by computing the weighted Euclidean distance between final segmentation results, where the weight is the length in base pairs of each segment. However, before I computed the Euclidean distance, I standardized the segmentation values by subtracting the mean and dividing by the standard deviation computed from the capture region log2 ratios. The variance correction significant decreased the weighted Euclidean distance between CNVkit and CopywriteR results (p-value < 0.01, paired Student t-test).

Tumors sequenced using the NimbleGen kit had no suitable reference due to poor quality tail DNA and subsequent low-quality sequencing results; and the livers could not be used as a control either due to being sequenced with a different capture kit. Instead, all tumors sequenced with the NimbleGen kit were used as a combined reference using CNVkit. After CNVkit corrected the combined reference for GC and mappability biases, I applied an additional correction that mitigated any bias introduced by any recurrent regions of amplification or deletion in the tumors using local regression (LOWESS). I expected there to be even coverage across the genome, which I confirmed by looking at the median-centered log2 coverage in the liver samples plotted along the genome. First, I sorted the probes by genomic position, and for each chromosome separately, applied our LOWESS correction using the *loess* R function with span=0.1 and where data was weighted by their normal density from the *dnorm* R function where *mean* was set to median and *sd* set to the standard deviation

across all log2 coverage values. Weighting the data was necessary to reduce the impact of outlier and low-coverage regions.

I used a previously published approach to identify thresholds for copy number changes in each sample (Geyer et al., 2017). Briefly, I sorted the log2 ratios of each capture region and calculated the median and standard deviation from the 50% central exon capture regions. I defined copy number segments with log2 ratio -2.5 or -7 standard deviations below the median as copy number losses and deep deletions, respectively. Copy number segments with log2 ratio +2 or +6 standard deviations above the median were copy number gains and amplifications, respectively. Finally, for the purpose of visualization and comparison across samples, I rescaled the log2 copy number ratios for each sample by dividing the negative log2 ratios by the deep deletion threshold and positive copy number ratios by the amplification ratio.

I found that tumors from our mouse model were moderately genomically unstable and had complex CNA landscapes (Figure 2-9). A heterogenous CNA landscape is in accordance with what has been observed in human TNBC. Furthermore, the CNA rate is higher in the *Brca1* deletion group (Figure 2-10), which was expected and has been previously been observed given the role *Brca1* plays in DNA repair (Holstege et al., 2010).

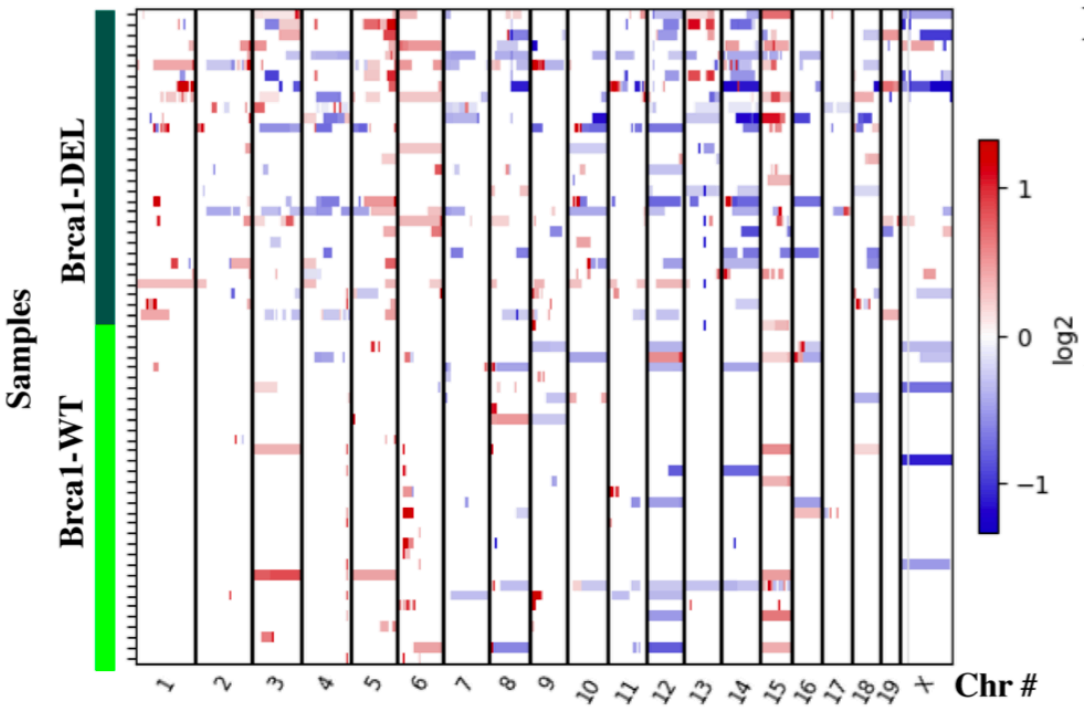

Figure 2-9: **Copy-number alteration landscape of mouse TNBC.**
Copy gains are in red and copy losses are in blue. X-axis shows chromosome number and y-axis shows individual samples annotated with *Brca1* genotype.

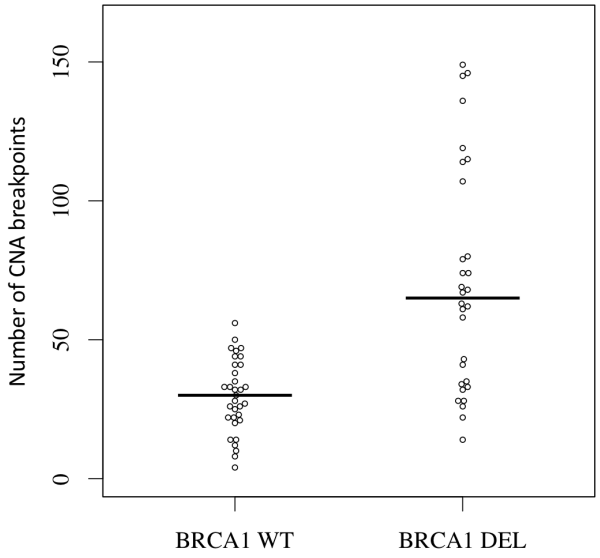

Figure 2-10: **CNA rate comparison between *Brca1* genotype.**
Number of CNA breakpoints per tumor is significantly different by *Brca1* genotype.

I found several known oncogenic genes that were subject to focal CNAs, including some that were recurrent. *Met*, *Myc*, and *Yap1* were focally amplified in the numerous samples (Figure 2-11). For most of the amplified samples, the expression of the amplified gene was also higher. Furthermore, I discovered several other known oncogenic genes that were amplified or deleted in single samples, including *Pten* deletion, *Fgfr2* amplification, and *Egfr* amplification (Figure 2-12). *MET* (Graveel et al., 2009; Ponzo et al., 2009), *EGFR* (Brandt *et al.*, 2000; Marozkina *et al.*, 2008; Masuda *et al.*, 2012), *YAP1* (S.-S. Chang *et al.*, 2017; Yu *et al.*, 2015; Zanconato *et al.*, 2016), *FGFR2* (Easton et al., 2007; Katoh, 2016), and *PTEN* (J. C. Liu *et al.*, 2014; Wang *et al.*, 2016) have all been previously shown to drive TNBC. In addition to their putative driver role, *Met*, *Fgfr2*, and *Egfr* encode protein kinases.

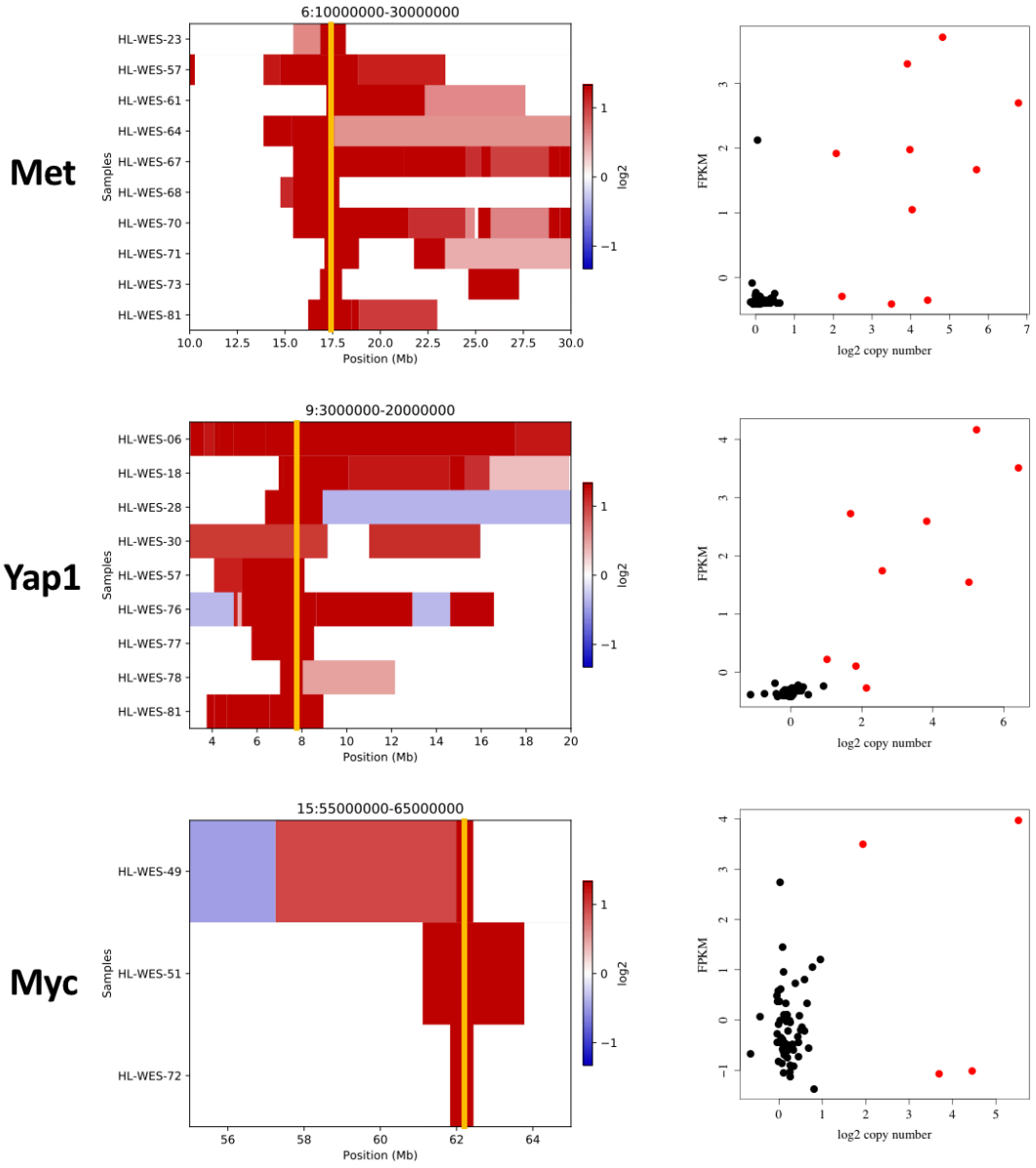

Figure 2-11: **Recurrent focal amplifications.**
Met (top row), *Yap1* (middle row), and *Myc* (bottom row) were focally and recurrently amplified in the mouse tumors. The left-hand side figures show the genomic position of the amplification for each sample (one per row). The vertical yellow line is the approximate location of the gene. The right-hand side figures plot the expression of each gene against the log2 copy number of the gene. Samples in red are those that were amplified based on the CNA data.

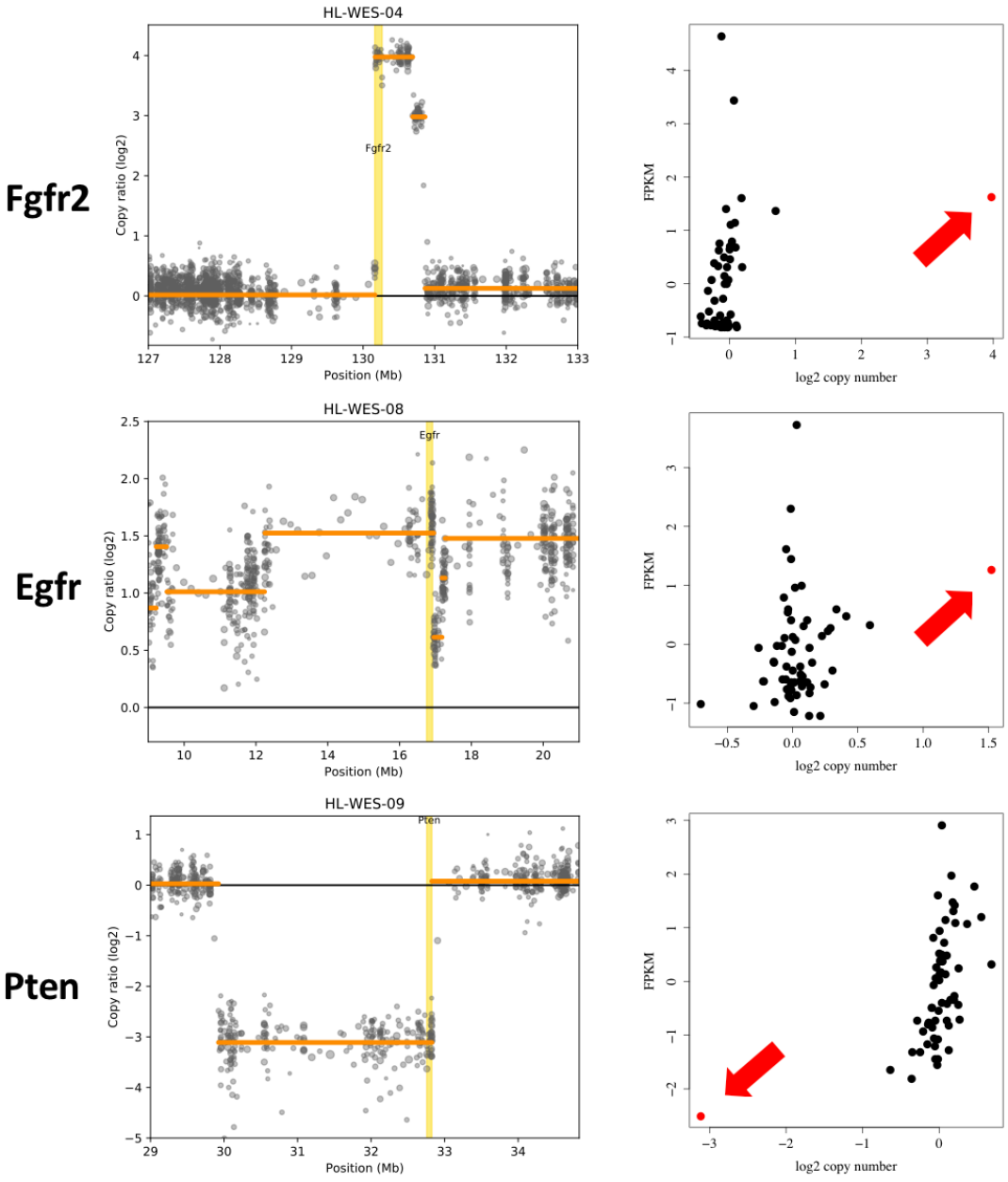

Figure 2-12: **Single focal amplifications.**
Fgfr2 (top row), *Egfr* (middle row), and *Pten* (bottom row) were focally amplified in single tumors. The left-hand side figures show the genomic position of the amplification for each sample (one per row). The vertical yellow line is the approximate location of the gene. The right-hand side figures plot the expression of each gene against the log2 copy number of the gene. Samples in red are those that were amplified based on the CNA data.

2.2.3 Mutations

Human breast cancer is known to have a relatively low mutation rate compare to other cancers, with an average rate of 90 mutations per tumor in the coding exons (Figure 2-13). Human TNBC is a little higher at 127 mutations per tumor. Furthermore, there are very few recurrent mutations in TNBC. One study found on 104 TNBC patients found *TP53* to be the most frequent (62%), followed by *PIK3CA* (10.2%), *USH2A* (9.2%), *MYO3A* (9.2%), and *PTEN* and *RB1* at 7.7% (Shah et al., 2012).

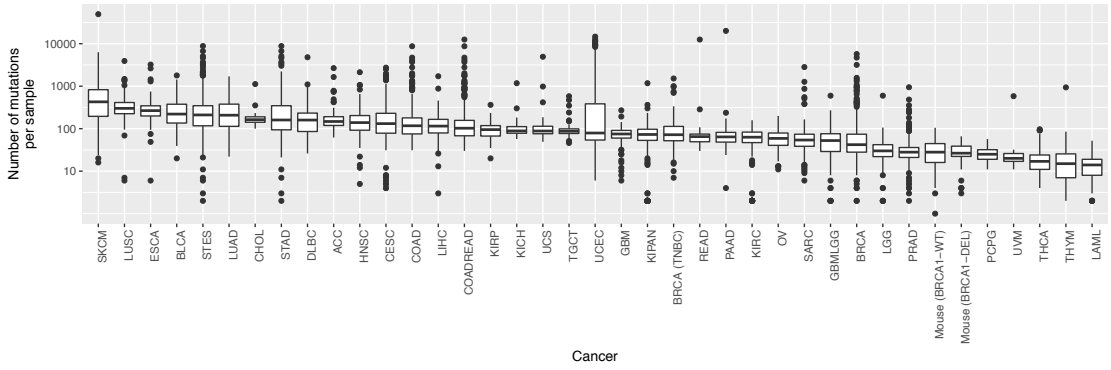

Figure 2-13: **Pan-cancer mutation rate.**
Plotted is the average number of mutations in the coding exons per tumor for various cancer types in TCGA.

I called mutations using Varscan2 on each tumor-normal pair (Koboldt, Zhang, *et al.*, 2012). If no normal tail sample was available for a tumor, I combined the three normal liver samples as a normal control for somatic mutation calling. For calling mutations from RNA-seq I applied the additional pre-processing step of removing PCR duplicate reads. Varscan2 default parameters were used except: minimum read

mapping quality of 15, minimum base quality of 15, at least four supporting reads, minimum sequencing depth of 12, minimum variant frequency of 5%, and p-value ≤ 0.01. Since not all sequenced tumors had matched normal controls, I applied a stringent filtering criterion to reduce potential false positive germline mutations. I removed mutations that had 4 or more variant supporting reads in any of the control samples (with minimum base and mapping qualities of 10) or for which there were fewer than 10 normal control samples with coverage less than 12. I then removed any mutations that were outside the exon capture regions, were found in any of the immunoglobulin, histocompatibility, or killer cell lectin like receptor genes were found in mouse dbSNP (Mouse Genomes Project Release Version 5) or were in known RNA-editing sites (Danecek et al., 2012). The database of known RNA-editing sites was downloaded (ftp://ftp-mouse.sanger.ac.uk/REL-1202-RNAEditing/RNA-editing.vcf.gz) then converted from mm9 to mm10 coordinates using CrossMap (Zhao *et al.*, 2014). I noticed hundreds of mutations that were specific to and shared among many of the RNA-seq samples and reasoned they were likely false positives. I filtered those mutations using the Fisher's exact test where the contingency table has the number of samples that have three or more variant supporting reads or less than three for the rows, and whether the sample is from RNA-seq or WEX for the columns. I removed mutations with p-value ≤ 0.01. Finally, I annotated the mutations using SnpEff (Cingolani et al., 2014).

I found an average of 30 mutations in the coding exons per mouse tumor, ranging from 0 to 104 (Figure 2-14). Compared to mutation rates within coding exons in human cancers, our mouse model has a relatively low rate (Figure 2-13). The two

most interesting mutations (Kras Q61H and Hras Q61H) overlapped with known hotspot mutations in cancer and are known drivers of cancer (http://cancerhotspots.org). I otherwise found few recurrent mutations, but found many mutations that occurred within protein kinases, many of which are known cancer drivers (Figure 2-15).

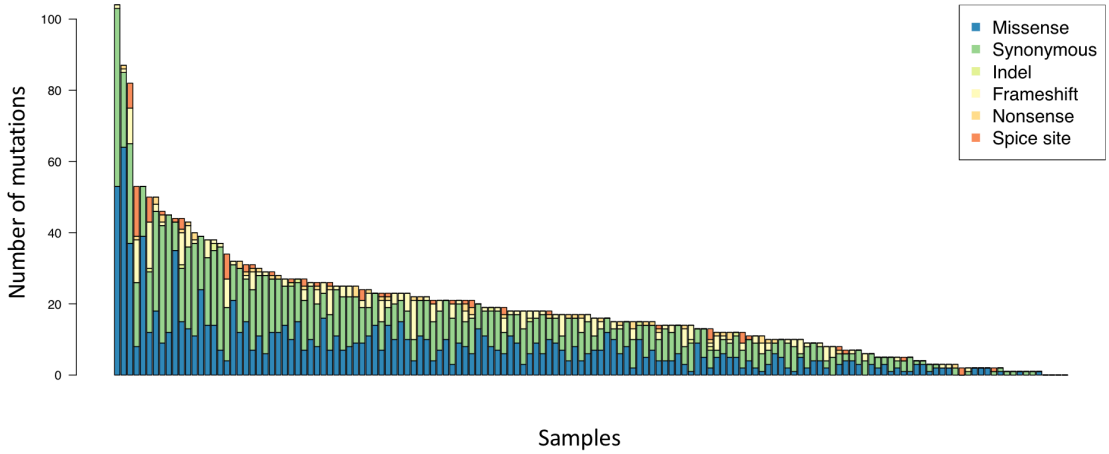

Figure 2-14: **Number and type of mutations in mouse TNBC.**
Shown is the number and breakdown by type of mutation for each mouse tumor.

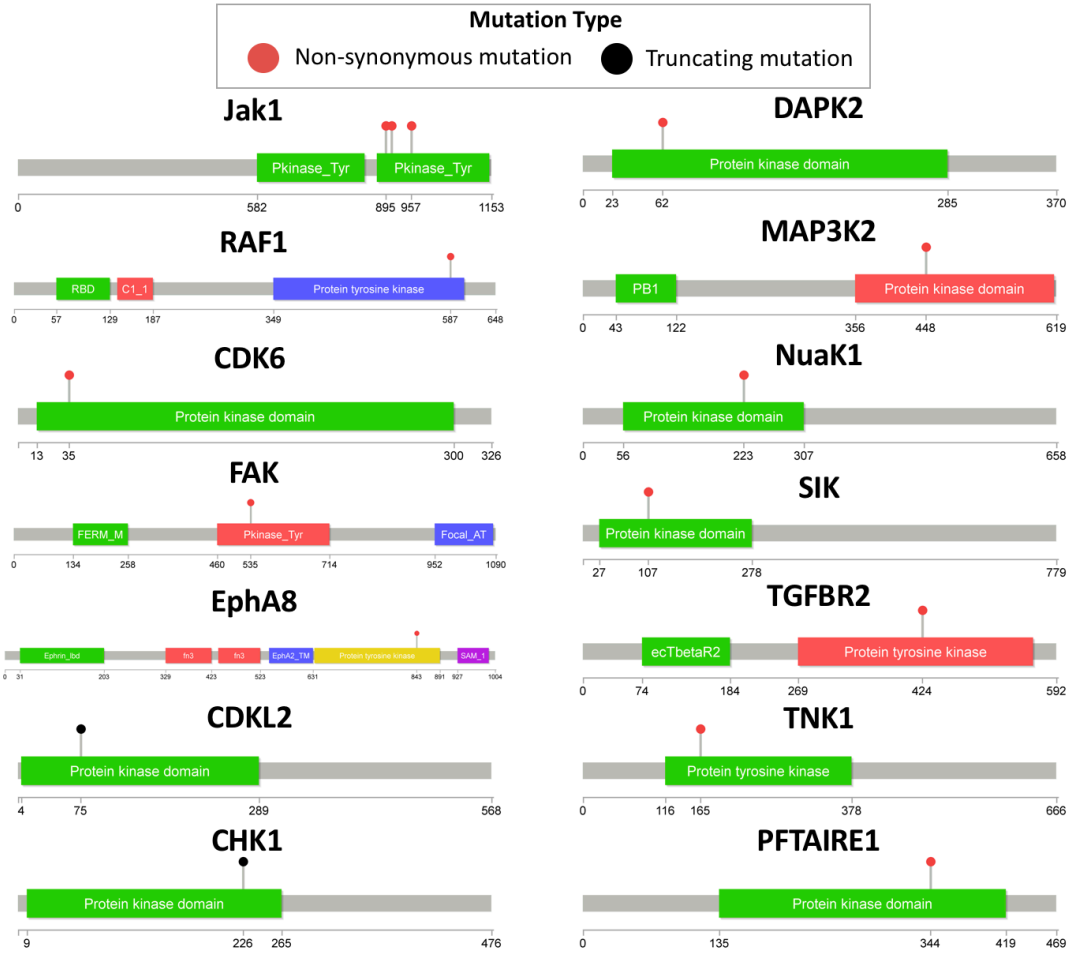

Figure 2-15: **Mutations in kinase domains**.
Lolliplot plots of mutations that are within kinase domains.

2.3 Mouse precision medicine identifies driver kinases

Precision medicine assumes that the cancer drivers are at least partially specific to each patient's tumor (Sboner and Elemento, 2016). Hence, the treatment given will be tailored to the germline and somatic alterations in each patient. Many research and organizational efforts are being made to deliver precision to cancer patients (Letai, 2017). Chantal *et al.* developed a pipeline using whole-exome sequencing and a living

biobank of patient-derived tumors for drug screens (Pauli et al., 2017). They applied their pipeline to 769 cancer patients from across a diverse cancer types and identified alterations in known cancer genes in 95.8% of patients. But there were only able to identify therapeutic vulnerabilities for 9.6% of patients based on their genetic data. They then derived tumor organoids for 56 of the samples and applied high-throughput drug dose-response screening on four of them to identify potential therapeutic vulnerabilities in the tumor. They were able to identify effect drug combinations for all four samples tested. Another precision medicine study is the MOSCATO 01 prospective clinical trial (Massard et al., 2017). The clinical trial recruited 1,035 patients with diverse cancer types and acquired a molecular portrait (with one or more of RNA-seq, aCGH, WES, and targeted sequencing assays) for 843 of them. They were able to identify an actionable molecular target in 49% of tumors (411/843).

Here I applied a precision medicine strategy with the sequencing data from our TNBC GEM model. I used gene expression, mutations, gene fusions, and CNAs along with publicly available databases to identify potential tumor drivers and the likely concomitant drugs that could be used to treat the tumor.

2.3.1 Identification of drivers and targetable alterations

I compared the somatic alterations found in our mouse tumors to the OncoKB database to identify oncogenic and actionable alterations, and the Cancer Hotspots database to identify mouse mutations which overlap with identified hotspots in human cancer (M. T. Chang *et al.*, 2016; Chakravarty *et al.*, 2017). I also supplemented data

from OncoDB with results from our own findings in the mouse tumors. I also identified which somatic alterations overlap genes in the PI3K and MAPK pathways (as defined by KEGG) to discern which tumors have alterations that potentially activate those pathways.

To identify overlapping point mutations, I first globally aligned the peptide sequences of canonical isoforms of mouse-human homologs using the Needleman-Wunsch algorithm with the BLOSUM62 matrix (Rice *et al.*, 2000). I then converted the amino acid coordinate of the mouse mutation to human coordinates. To identify overlapping CNAs, I defined deleted mouse genes as those with copy number 0 and amplified genes as those with copy number greater than 3. Finally, I used a z-score threshold of +/- 1.5 to identify significant over or under-expressed genes.

I found oncogenic alterations in 76.7% (56 of 73 tumors) of tumors and targetable alterations in 49.3% (36 of 73 tumors) of tumors (Figure 2-16). I also found that 75% (54 of 72) of tumors harbored at least one somatic alteration (CNA, mutation, or fusion) in any of the mouse protein kinases. Fifteen kinases where affected by two or more somatic alterations (Figure 2-17).

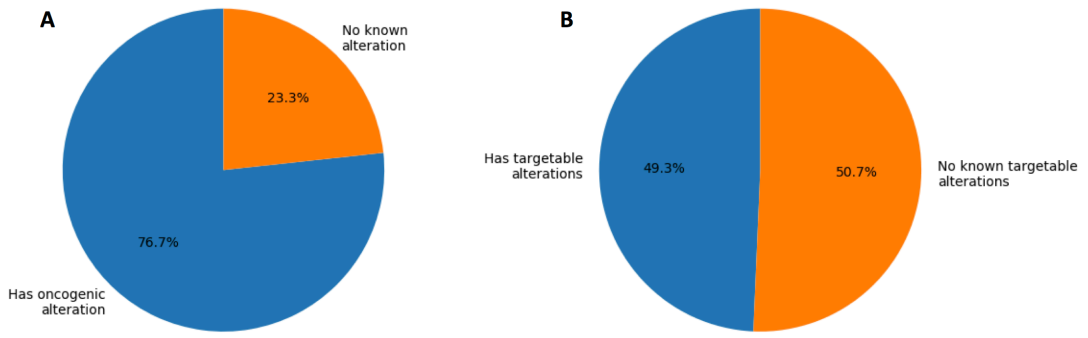

Figure 2-16: **Oncogenic and targetable alterations.**
(A) Shows the proportion of tumors that were identified as having a known oncogenic alteration. (B) Shows the proportion of tumors with alterations that can be targeted by existing drugs.

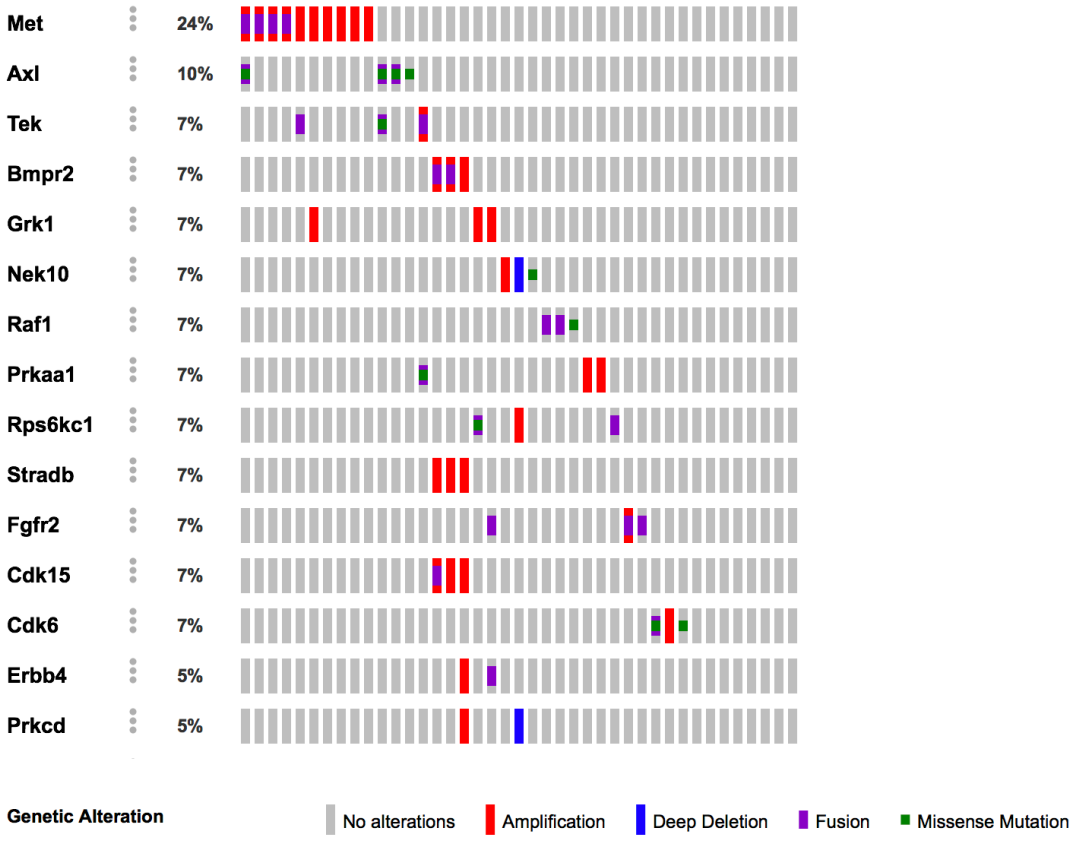

Figure 2-17: **Kinases with two or more alterations.**
Rows are individual kinases that were affected by two or more somatic alterations. Columns are the primary tumors.

2.3.2 Precision medicine treatment

Our next step was to take the identified oncogenic alterations and apply treatment to a select number of mice. Based on our *in vitro* validation for the presence of the alterations, we chose to target tumors containing *Fgfr2* fusions, *Raf1* fusion, *Braf* fusion and *Met* overexpression. For each target, we chose drugs that were (i) have high specificity and high competence indicated by low $IC_{50,}$ (ii) have good *in vivo* bioavailability with low *in vivo* dosing, and (iii) are approved for cancer treatment or are currently in phase II or III clinical trial.

We treated tumors with either Fgfr2-Dnm3 or Fgfr2-Tns1 gene fusions with the FGFR inhibitor, BGJ398. We first validated that the fusion is present using sanger sequencing on the cDNA (Figure 2-18). We also demonstrated that the tumors with the fusion had higher levels of phospho-FSR2, which is an indicator of FGFR pathway activation (Figure 2-19). We treated both Fgfr2 fusions and found that BGJ398 alone was sufficient to lead to complete tumor regression (Figure 2-20). We also included a couple of negative control treatments. In addition to treatment with no drug, we used Met inhibitor, crizotinib, on the Fgfr2-Dnm3 since the tumor has no Met alterations. As expected, in all cases the control treatment did not halt tumor growth. Furthermore, we applied combination treatments to both Fgfr2 fusions. The tumor with Fgfr2-Dnm3 has the *Brca1* deleted genotype, so is sensitive to the PARP inhibitor, Olaparib. Combining Olaparib with BGJ398 lead to complete tumor remission, as was the case for BGJ398 alone, but also prevented any tumor relapses for 80 days; whereas BGJ398-alone treatment had a 50% recurrence rate (data not shown). For the tumor

with Fgfr2-Tns1 fusion, we applied the PI3K inhibitor, BKM120. By itself it slowed tumor regression, but when combined with BGJ398 they worked synergistically to produce fast and complete tumor remission.

We treated the tumor with the Dhx9-Raf1 gene fusion using the MEK inhibitor, trametinib (GSK1120212). We validated the presence of the fusion using sanger sequencing on the fusion cDNA (Figure 2-18). We also demonstrated that tumors with the fusion had higher phospho-ERK levels, which is an indicator of MAPK pathway activity (Figure 2-19). We then treated the tumor with trametinib and found that it alone significantly delayed tumor growth (Figure 2-20). Moreover, since the tumor with the Dhx9-Raf1 fusion was also *Brca1* deleted, we treated the tumor with olaparib alone or in combination with trametinib. Olaparib alone resulted in slowly causing tumor remission, but when combined with trametinib it led to complete tumor remission.

We also treated the tumor with the Dlg1-Braf gene fusion using trametinib. After validating the presence of the fusion with sanger sequencing (Figure 2-18), we showed that the tumors with the fusion had higher phospho-ERK signaling (Figure 2-19). Hence, we reasoned that since the tumors had higher MAPK signaling they may also be sensitive to MEK inhibition using trametinib. However, treatment with trametinib alone only slightly slowed tumor growth (Figure 2-20). Moreover, combined trametinib and BKM120 performed slightly better than BKM120 alone, but still did not lead to tumor remission.

Finally, we treated a tumor overexpressing Met with the Met inhibitor, crizotinib. We validated the presence of the fusion by sanger sequencing (Figure 2-18)

and showed that tumors with the fusions had higher phospho-ERK (Figure 2-19). Treatment of the Met-overexpressed tumor with crizotinib alone only slightly delayed tumor growth (Figure 2-20). Similarly, treatment with BKM120 alone also only delayed tumor growth. However, combined crizotinib and BKM120 treatment lead to tumor regression.

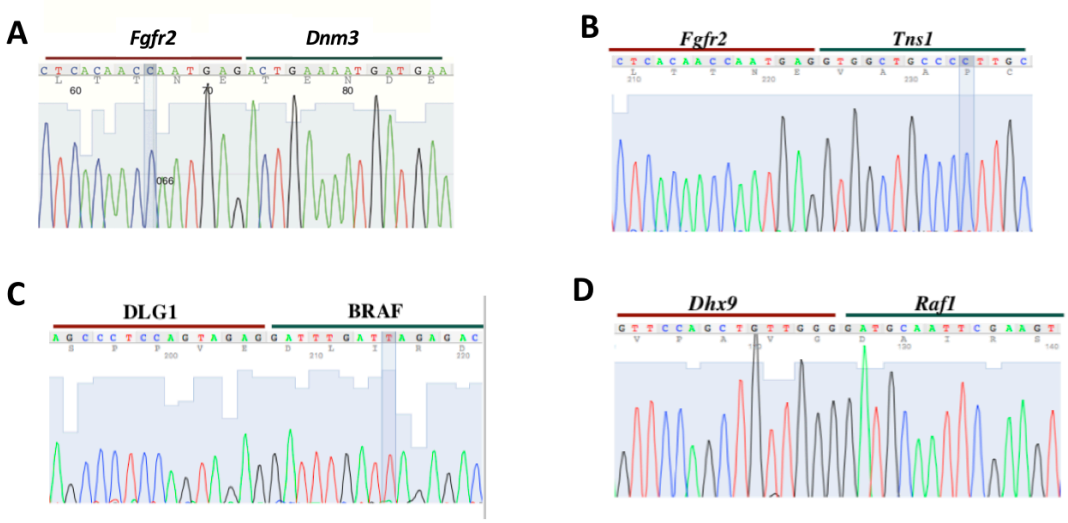

Figure 2-18: **Sanger sequence validation.**
(A) Fgfr2-Dnm3 fusion. (B) Fgfr2-Tns1 fusion. (C) Dlg1-Braf fusion. (D) Dhx9-Raf1 fusion.

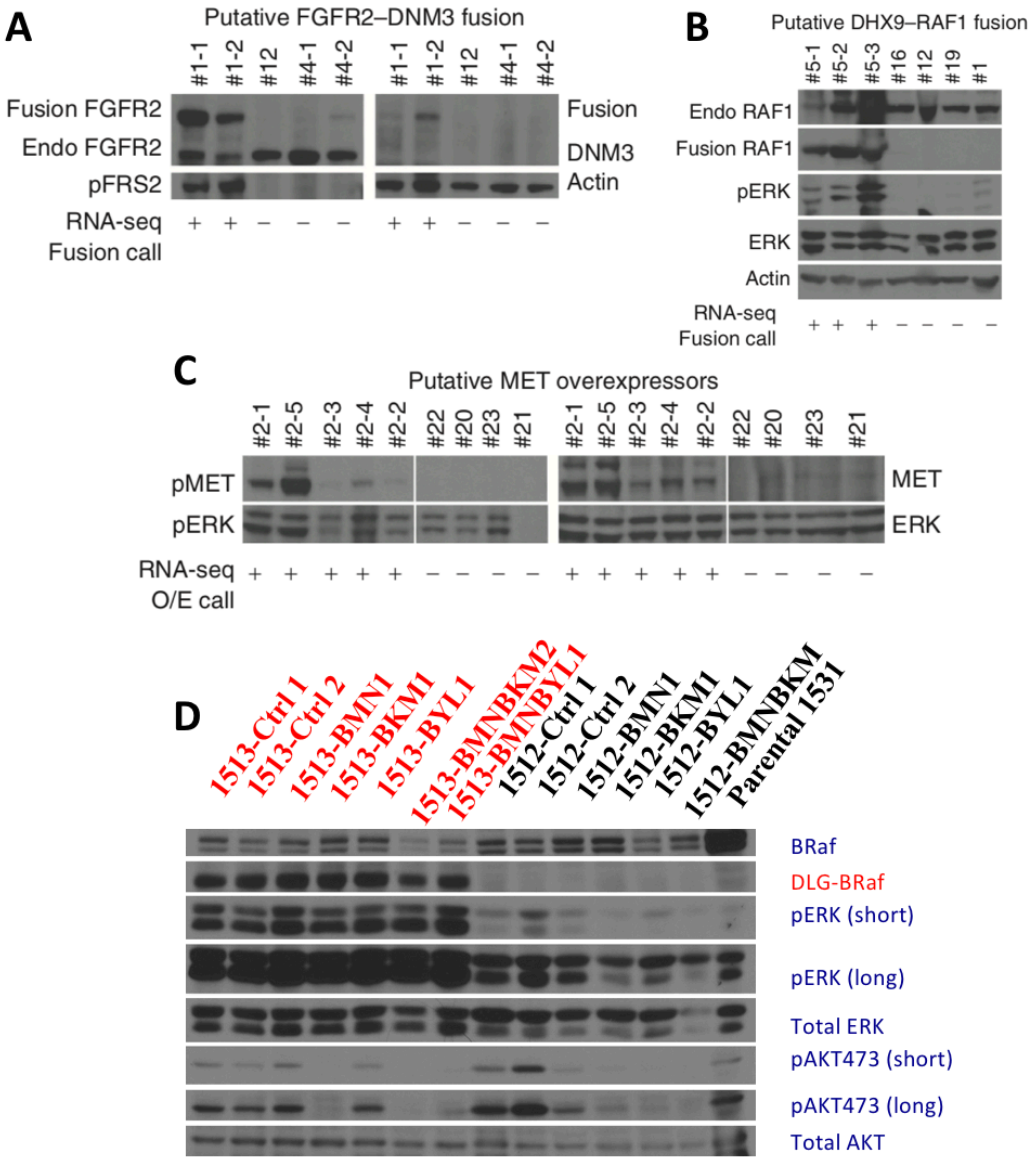

Figure 2-19: **Western blots.**
(A) Western blot showing increased pFRS2 levels in tumors with the Fgfr2-Dnm3 fusion. (B) Western blot showing increased pERK levels in tumors with the Dhx9-Raf1 fusion. (C) Western blot showing increased pERK levels in tumors with the Met over-expression. (D) Western blot showing increase pERK levels in tumors with the Dlg1-Braf fusion.

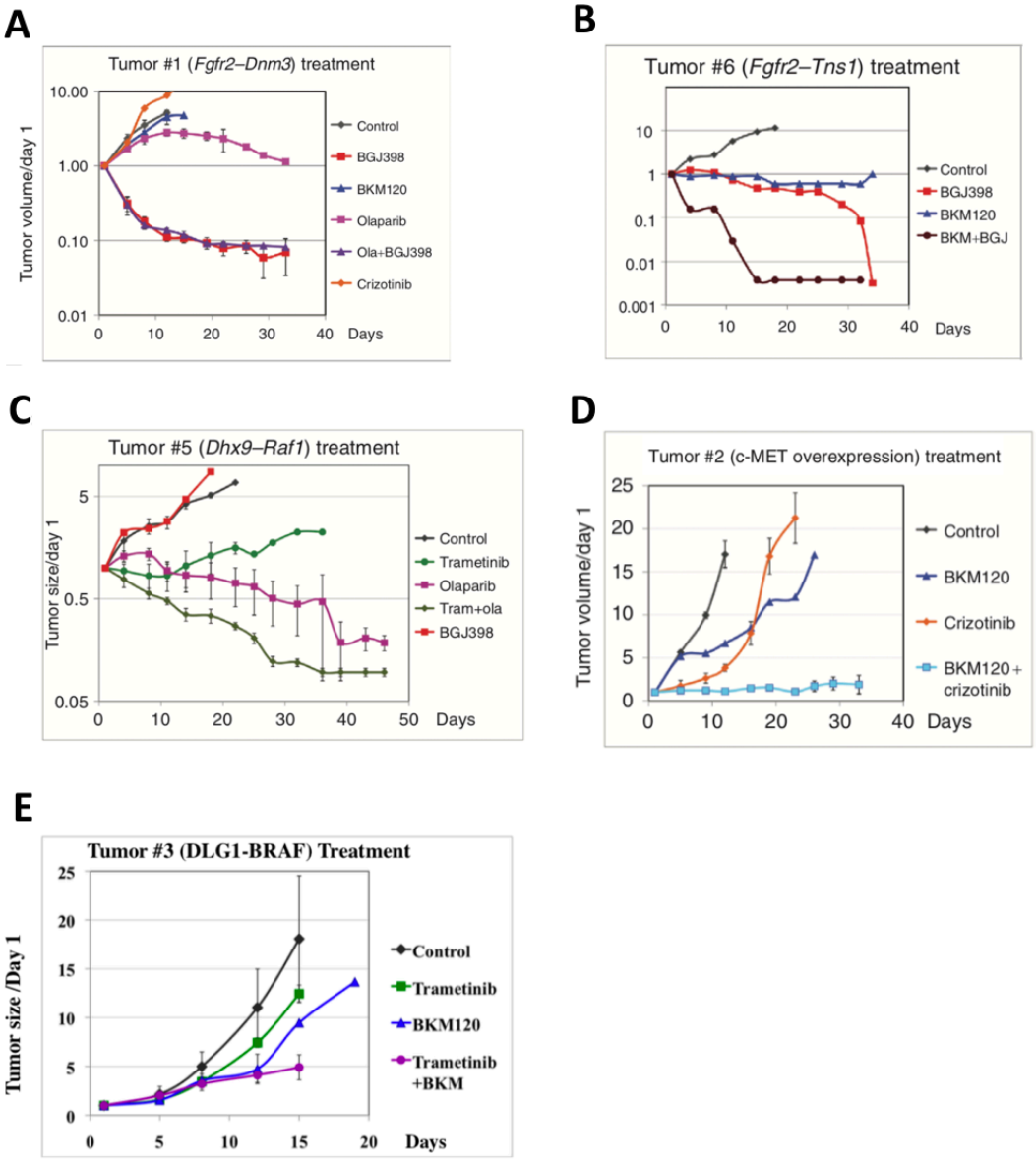

Figure 2-20: **Tumor treatment.**
(A) Treatment results on tumors with the Fgfr2-Dnm3 fusion. (B) Treatment results on tumors with the Fgfr2-Tns1 fusion. (C) Treatment results on tumors with the Dhx9-Raf1 fusion. (D) Treatment results on tumors with Met over-expression. (E) Treatment results on tumors with the Dlg1-Braf fusion.

In summary, we treated five different somatic alterations involving different protein kinases. We were able to validate the presence and signaling consequences for all alterations. Most treatments were successful when targeting the kinase alone or in combination of another drug. The FGFR inhibitor, BGJ398, was exceptionally effective alone in causing tumor remission in Fgfr2-fused tumors. We also found that inhibitors targeting Raf1 and Met in Dhx9-Raf1 and Met-overexpressed tumors, respectively, were significant in delaying tumor growth. Moreover, when combined with other inhibitors they were even more effective than single-target therapy. Finally, we applied a MET inhibitor to the tumor harboring the Dlg1-Braf fusion, but only saw a slight delay in tumor growth.

2.4 Summary

In summary, I profiled 72 primary TNBC tumors that were generated from *Trp53*-deficient, with or without *Brca1* deficiency, mice. I used a combined RNA-seq and WES approach to verify that the GEM model accurately reflects human TNBC, identified somatic alterations, and applied treatment using a precision medicine approach. Moreover, the analyses revealed a large number of protein kinases that were somatically altered, some of which were validated to be driving tumor growth. Our results indicate the validity of a precision medicine approach for treating mouse TNBC as well as identifying protein kinases that driver tumor growth.

The first step in my analysis was to validate that the GEM model reflects human TNBC. I verified by IHC and RNA-seq that most tumors lack expression for

PR, ER, and HER2, which is what defines human TNBC. Moreover, I used two gene expression classifiers to demonstrate that most mouse tumors are of the basal-like subtype, which is the breast cancer subtype most often associated with TNBC. Finally, I found that the mouse tumors exhibited low mutation rates and high rates of gene fusion and CNAs, which are all features of human TNBC.

I next sought to characterize the genomic and transcriptomic landscape of the mouse tumors. Our first finding revealed a large number of heterogenous gene fusions. Few fusions were recurrent across the tumors but there were a number of genes that were fused in multiple tumors but involved different gene partners. For example, I found *Fgfr2* fused to three different gene partners: *Tns1*, *Dnm3*, and *Zmynd8*. I also found number of other gene fusions involving protein kinases, including: Dhx9-Raf1, Rpl32-Raf1, Dlg1-Braf, Met-Calu, and Met-Itsn1. Next, I surveyed the mutational landscape and found low mutation rates and no recurrent mutations. However, I did find two mutations that occurred in known hotspot mutations in human cancer: KRAS Q61H and HRAS Q61H. I also found a number of nonsynonymous and truncating mutations in protein kinases. Finally, I surveyed the CNA landscape and found a very heterogenous number of CNAs. I found three recurrent CNAs: *Met* (a protein kinase), *Yap1*, and *Myc*. I also found focal CNAs in single tumors affecting known oncogenic genes, including: *Pten*, *Egfr*, and *Fgfr2*.

I applied a precision medicine approach to the mouse tumors and selected a number of somatic alterations for follow-up validation and treatment. I chose the following: Fgfr2-Dnm3, Fgfr2-Tns1, Dlg1-Braf, Dhx9-Raf1, and Met over expression. Targeted therapy of *Fgfr2* fusions using FGFR inhibitors was very successful at

causing tumor remission. Targeting Dhx9-Raf1 fusion and *Met* over expression with therapies significantly delayed tumor growth, but when combined with other therapies (BKM120 or Olaparib) lead to tumor remission. We attempted to target the Dlg1-Braf fusion alone with a MEK inhibitor, but only achieved a minor delay in tumor growth. Moreover, combining the MEK inhibitor with BKM120 increased efficacy, but still only lead to delay in tumor growth and not tumor remission.

CHAPTER 3

KINOME SUBSTRATE SPECIFICITY

Understanding which kinases phosphorylate which substrate proteins is of vital importance for understanding cellular signaling networks. Kinases use several contextual cues that determine which substrates it can phosphorylate, including amino acid composition around the phosphorylated serine, threonine, and tyrosine, tissue and cellular specificity, and protein interaction networks.

Over the past few decades, several categories of biochemical and molecular technologies have been created to determine the substrates of a given kinase, but each has its own set of limitations. Here I outline the general types of technologies, but the reader can refer to several good reviews for more in depth discussion (de Oliveira et al., 2016; Newman et al., 2014). Some technologies use radio-labeled ATP substrate incubated with the kinase and protein substrate of interest to determine reactivity, but the main limitation of this approach is that it is low-throughput and many protein kinases will phosphorylate almost any protein if incubated long enough, thereby resulting in false positives. Protein microarrays print thousands of purified substrate proteins to an array and are then incubated with a solution containing ATP and the kinase-of-interest. However, since the reaction mixture occurs outside of the physiological conditions of the cell, the identified phosphorylation event may not really occur in vivo. Additionally, if the phosphorylation event depends on scaffolding proteins or other protein interactions, protein microarrays may miss them. The peptide

library (discussed later) is another approach with similar limitations to protein microarrays. However, one difference is the protein microarray cannot tell where on the protein it is being phosphorylated, whereas the peptide library can tell you the amino acid context each kinase prefers. Other approaches make use of mass spectrometry (MS) for measuring changes in phosphorylation levels of substrates after genetic or pharmacological modification of the signaling pathway. In these approaches the peptide mixture prior to MS can be phospho-peptide enriched or kinase-specific antibodies can be used while the kinase is bound to the substrate. However, since kinase signaling networks are complicated and phosphorylation can be transient, the identified phosphorylation events cannot always be detected or reliably attributed to the kinase-of-interest.

Computational approaches are an alternative approach for determining kinase substrate specificity, with over 50 computational tools that predict kinase substrate relations currently published (Bórquez and González-Billault, 2016; de Oliveira *et al.*, 2016). However, these approaches also have some important limitations. Firstly, the vast majority of these tools depend on databases of reported kinase-substrate pairs determined from literature and mass-spectrometry experiments, such as PhosphoSitePlus and Phospho.ELM (Hornbeck *et al.*, 2004; Dinkel *et al.*, 2011). There is the potential that many of the reported kinase-substrate pairs may be false. That is, some may not actually occur in vivo or the wrong kinase was attributed as being responsible for the site. Secondly, although there are hundreds of thousands of reported phosphorylation sites, the vast majority do not have a known kinase assigned to them. In addition, there is wide variability in the number of substrates reported for

each kinase and most kinases have few or no known substrates. Hence, training a classifier for most kinases is not feasible, and for those which training data does exist there is a significant class size imbalance problem.

Computational and molecular approaches for determining kinase-substrate pairs both have their limitations, but the peptide library mentioned above provides a good compromise. The peptide library determines the kinase's amino acid preferences around the phosphorylated serine, threonine, and tyrosine in the C-terminal and N-terminal directions (Songyang *et al.*, 1994; 1995). The assay works by mixing the purified kinase of interest with a mixture of degenerate peptides and radiolabeled ATP. There are 198 degenerate peptide mixtures, where each contains peptides with a central, un-phosphorylated serine, threonine, or tyrosine, a fixed amino acid at one of the surrounding positions, and the remaining positions contain degenerate amino acids (equimolar amounts of each amino acid) (Figure 3-1). After the reaction has completed, aliquots of each are transferred to an avidin-coated membrane for imaging. Hence, the peptide library does not determine individual kinase-substrate pairs but indicates the kinase's preferences for the surrounding amino acid composition. Additional computational methods are necessary to take the kinase's specificity matrix with an appropriate scoring model and other contextual factors to determine which substrates the kinase can phosphorylate.

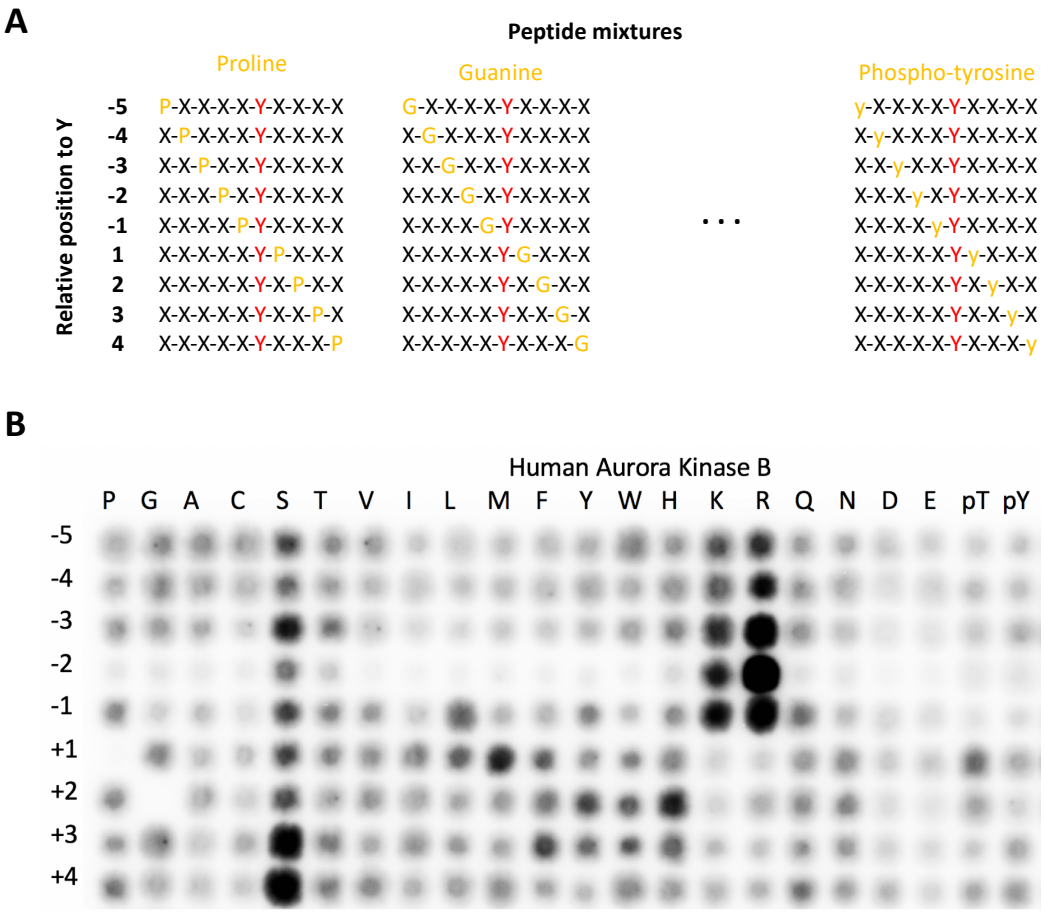

Figure 3-1: **Kinase peptide library.**
(A) Peptide mixture design for each well for measuring a tyrosine kinase's amino acid selectivity. The yellow character is the fixed amino acid that varies by position on the y-axis and amino acid type on the x-axis. The black X's are degenerate amino acids. The red "Y" is the central tyrosine that will be phosphorylated. (B) Example peptide library result for human Aurora Kinase B.

In this chapter I present peptide library data for several hundred human protein kinases and downstream computational analyses. Some parts of this analysis were performed in collaboration with Jared Johnson, who generated the peptide library data. I start with several analyses that validate the quality of the data and then demonstrating how I arrived at a scoring model that uses the peptide library for predicting kinase substrates. I then explore phospho-priming, which is a kinase signaling mechanism previously unstudied at the global level using computational methods.

3.1 Modeling kinase substrate specificity

The peptide library was performed on 262 kinases. My first task was to assess the quality of the data. I performed an unsupervised clustering (t-SNE) on the kinases (Figure 3-2 and Figure 3-3) (Maaten and Hinton, 2008), which is a machine learning algorithm to visualize high-dimensional data. t-SNE revealed that the kinases separate primarily by kinase type, with tyrosine kinases and serine/threonine kinases forming separate clusters. I also found through t-SNE that kinases generally group by family (Manning et al., 2002). Hence, the kinases cluster based on prior knowledge of kinase function and evolutionary history.

Next, I explored the relationship between kinase substrate specificity from the peptide library with the kinase domain sequences. All protein kinases share an evolutionarily conserved ~250 amino acid kinase domain. A phylogenetic tree based on the kinase domain has previously been used to reveal the evolutionary relationship

among the kinases and family groupings (Manning et al., 2002). However, not every amino acid in the kinase domain is important for determining kinase substrate specificity. Pau *et al.* developed a method to identify amino acids in the multiple sequence alignment of kinase domains that are more important for determining substrate specificity, the so called determinants of specificity (DoS) (Creixell, Palmeri, *et al.*, 2015). Furthermore, they found that the correlation between kinase substrate specificity and kinase domain sequence increased when restricting the domain to only the DoS. I performed a similar analysis using the 68 DoS with KINspect scores ≥ 0.9 and found a similar increased correlation between kinase domain and kinase specificity when restricting to only the DoS. (Figure 3-4).

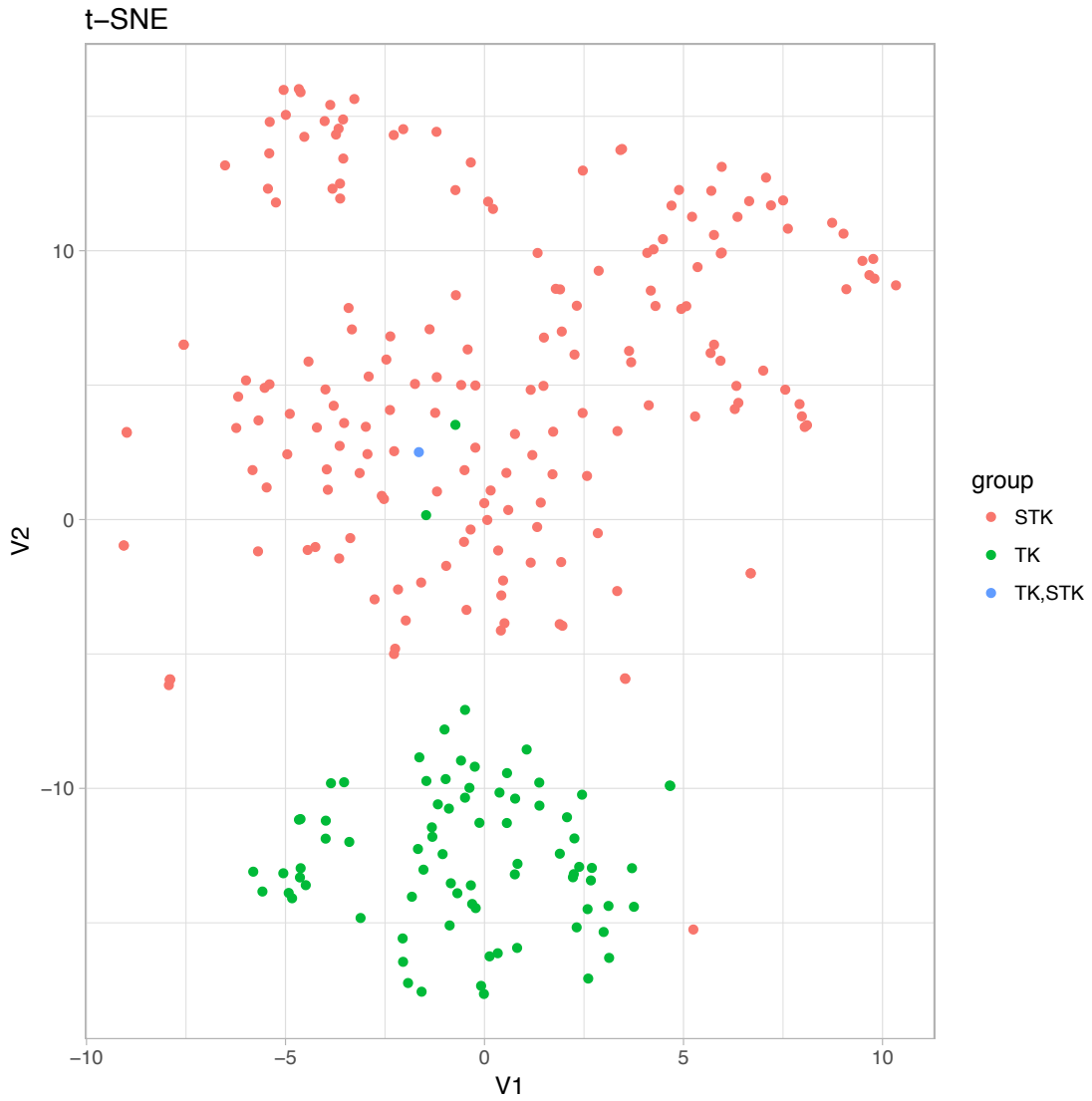

Figure 3-2: **t-SNE by kinase type.**
Two-dimensional t-SNE plot showing clear separation of the kinase by type (tyrosine kinases and serine/threonine kinases).

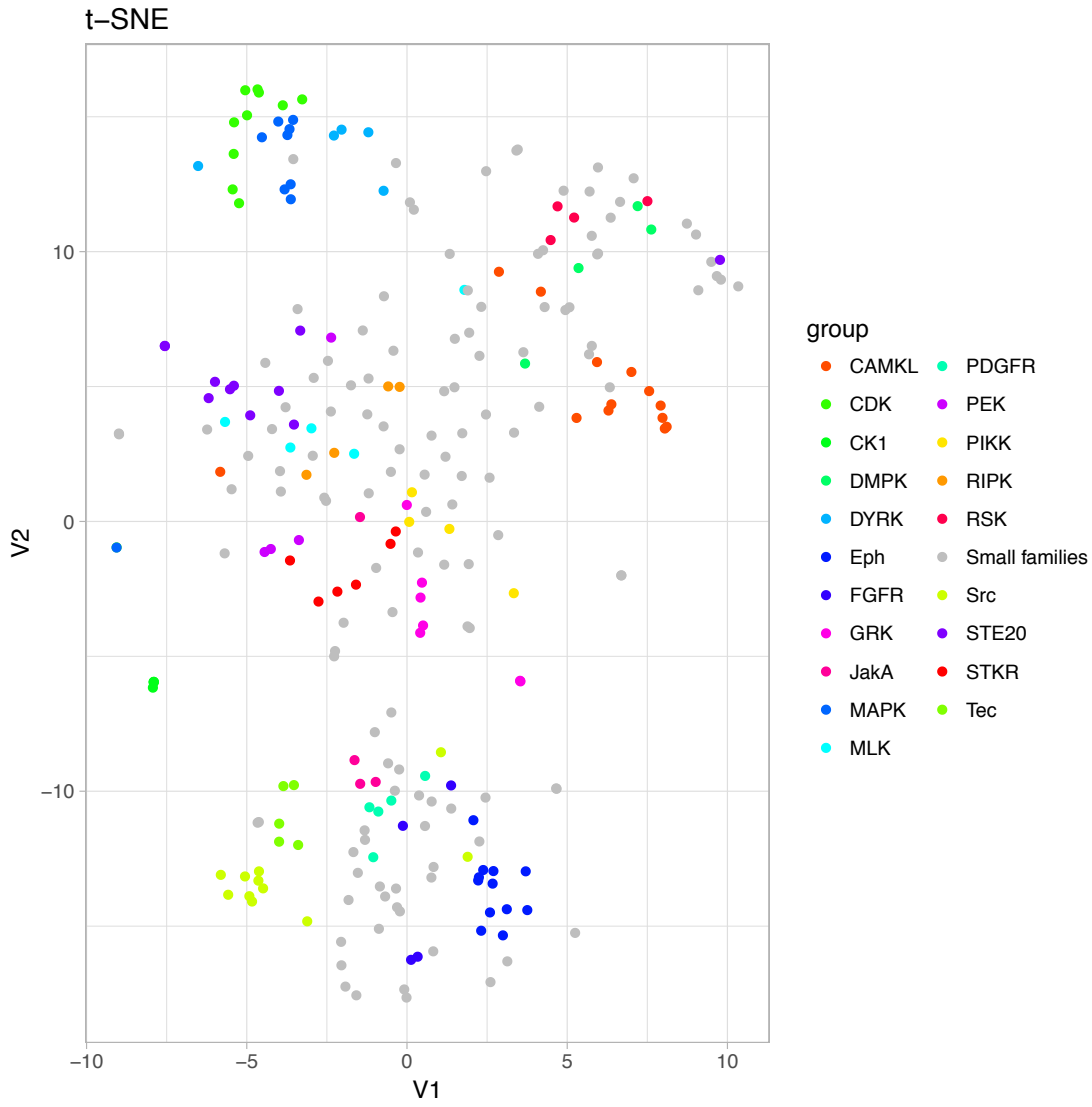

Figure 3-3: **t-SNE by kinase family.**
Two-dimensional t-SNE that shows clear grouping of kinases by kinase family.

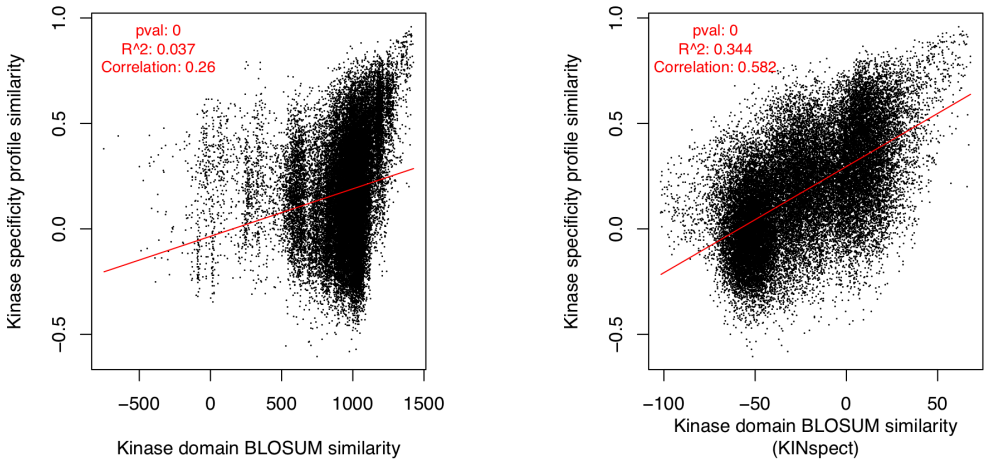

Figure 3-4: **Kinase domain and specificity correlation.**
Kinase domain BLOSUM similarity were computed from the multiple alignment and BLOSUM distance scoring approach used by Pau *et al.* Kinase specificity profile similarity was computed using pairwise distance between each peptide library profile with spearman correlation.

My next step was to benchmark the ability of the peptide library data to identify each kinase's known substrates. Several databases exist that curate the literature for published kinase-substrate relationships, such as PhosphoSite and Phospho.ELM (Hornbeck *et al.*, 2004; Dinkel *et al.*, 2011). Using data from PhosphoSite (version MAR_21_2017), I plotted the number of 'putative' substrates per kinase to demonstrate that most kinases have few or no reported substrates (Figure 3-5). Moreover, there are 233,567 listed phosphorylation sites, but only 7,141 of them assigned to one or more kinase. Therefore, training a classifier for most kinases is not possible, and even for those with hundreds of reported substrates still have a large class imbalance problem. Hence, making the task of predicting kinase substrates or benchmarking an existing algorithm to be challenging.

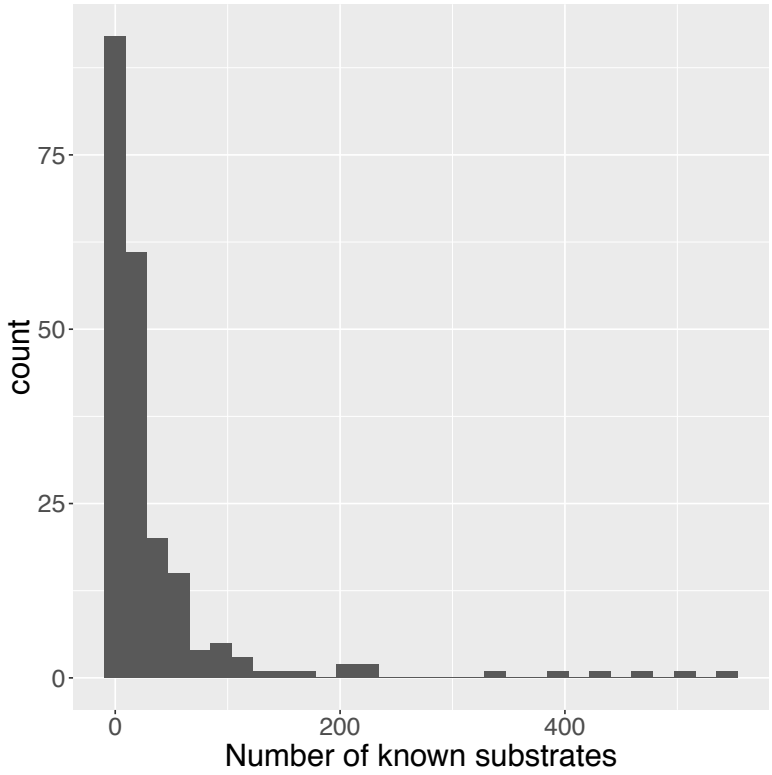

Figure 3-5: **Putative kinase substrates.**
Plot of the number of putative kinase substrates per kinase as listed in PhosphoSite.

I next attempted to filter 'putative' kinase substrates in PhosphoSite since false positive sites would lead the underestimating the performance of the peptide library. As explained earlier, some of the sites may be incorrectly reported by either being a phosphorylation site that does not actually occur or the wrong kinase was assigned to it. For each kinase for which peptide library data was available and had at least 10 reported substrates in PhosphoSite, I scored each and plotted against three metrics that may serve as indicators for the quality of the site. 'LT_LIT' is the number of publications supporting the phosphorylation site. 'MS_LIT' is the number of mass spec studies supporting the site. 'MS_CST' is the number of mass spec studies

performed by Cell Signaling Technology that support the site. I assumed that the higher any of the three metrics are the better quality the site. However, plotting LT_LIT, MS_LIT, and MS_CST against kinase score yields no obvious relationship (Figure 3-6). I also performed similar plots but for each kinase separately and got a few kinases with significant correlations, though most were negatively correlated (data not shown). Hence, little can be concluded from this analysis.

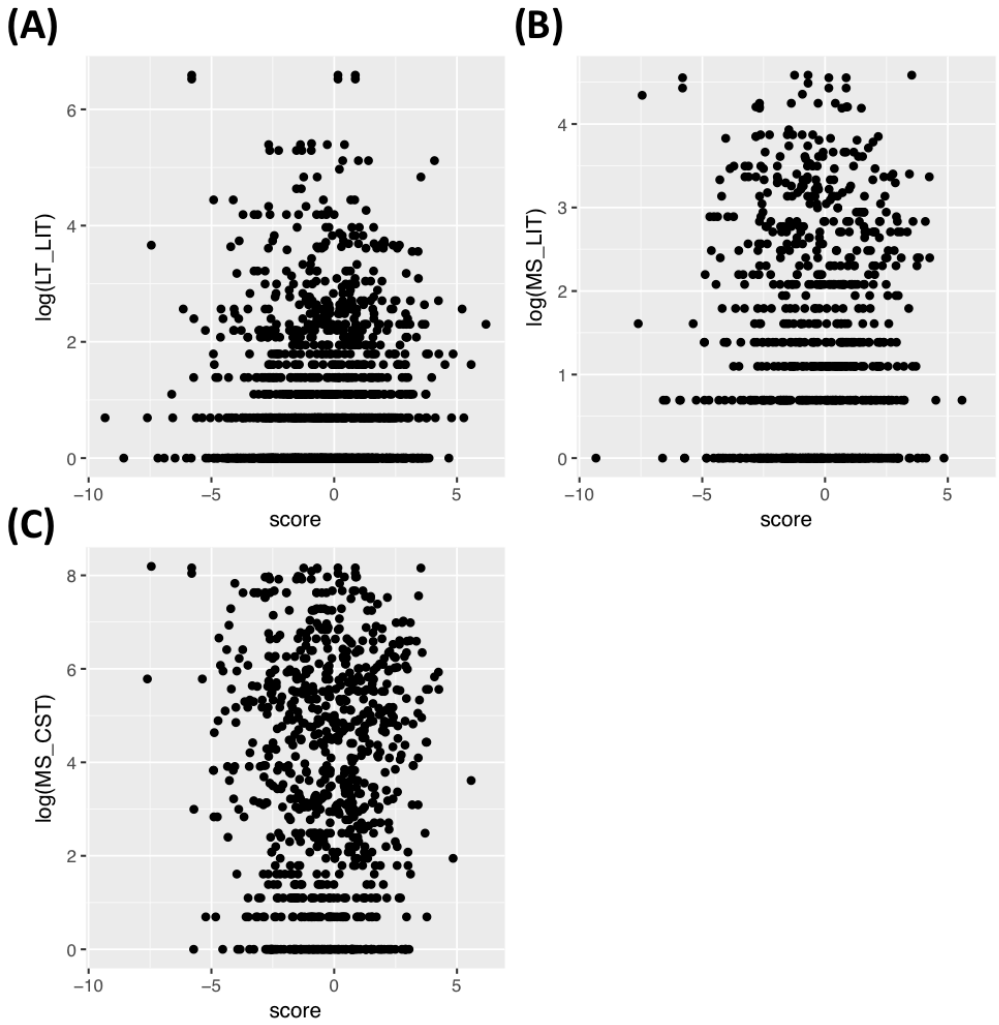

Figure 3-6: **Correlation with quality metrics.**
Plots of (A) LT_LIT, (B) MS_LIT, and (C) MS_CST measures of each phosphorylation site against the peptide library score from the kinase that is reported to phosphorylate it.

I then benchmarked the kinase specificity profiles from the peptide library to identify their reported substrates. I only used kinases that had at least 10 known substrates and applied a variety of normalization methods to the raw peptide library densitometry data. The scoring algorithm for each individual kinase works by computing empirical p-values for each candidate substrate to decide if the kinase can phosphorylate it or not. Computing the p-value for a candidate substrate works as follows: (1) use the kinase's normalized peptide library result to score all phosphorylation sites in PhosphoSite (tyrosine kinases are restricted to pY sites and serine/threonine kinases are restricted to pS/pT), and then (2) use the empirical distribution from step (1) to compute the p-value for the candidate substrate. I computed several performance metrics for kinase (Figure 3-7). The results show that serine/threonine kinases generally perform better than tyrosine kinases and that "row_sum" normalization works best. Overall, the peptide library does better than random, however, there is room for improvement since many of the published algorithms for predicting kinase substrates do significantly better.

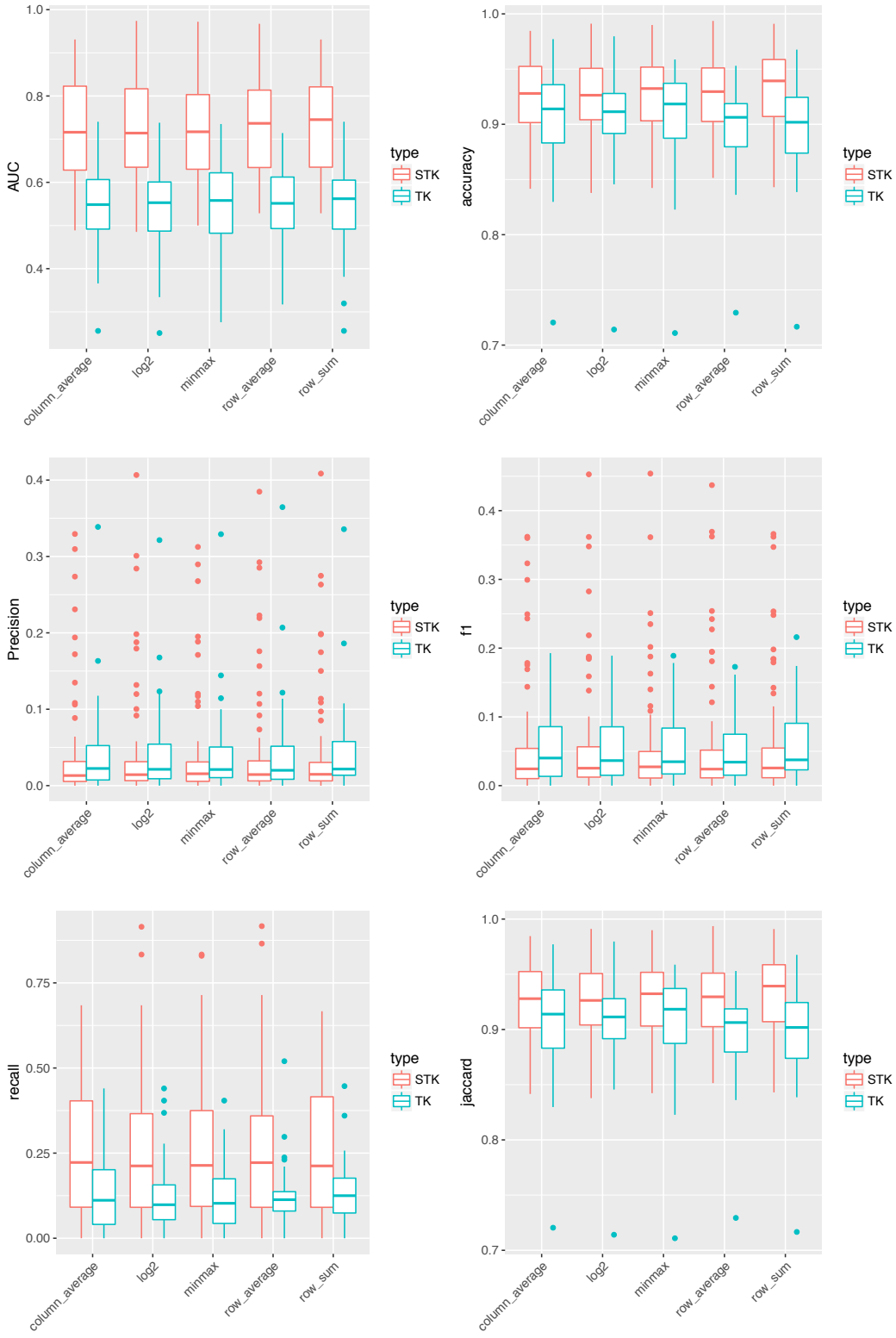

Figure 3-7: **Kinase peptide library performance.**
Metrics for the ability of the kinase peptide library to identify each kinase's reported substrates. The graphs are broken down by kinase type any normalization approach.

3.2 Phospho-priming

Many kinases depend on the presence of surrounding phosphorylated serines, threonines, or tyrosines before they can phosphorylate their substrate; a process referred to as phospho-priming. Previous research has demonstrated its role in phosphorylation signaling. The review by Valk *et al*. suggests that phospho-priming can permit more complex signaling mechanisms than are possible with singly phosphorylated sites (Valk et al., 2014). Some examples include: multiple phosphorylation sites can create a graduated input (switch-like behavior) and sequential phosphorylation sites that each require a different kinase can create an 'and' gate.

The peptide library data presented in this thesis measures kinase preference for pT and pY and can form the basis for further understanding phospho-priming at a global level. As an example, Figure 3-8 shows the pY preferences across the tyrosine kinases. Furthermore, to explore the potential prevalence of phospho-priming I examined the relative frequencies of different amino acids including pY, pS, and pT around reported phosphorylation sites (Figure 3-9). The figures demonstrate that for all three types of phosphorylation sites, there is a significant enrichment of other phosphorylation sites. Indicating that there is the potential for phospho-priming to occur. However, one alternative explanation for the co-occurrence of phosphorylation sites is intrinsic instability (He et al., 2009). Portions of many proteins have stretches of amino acids that do not form secondary or tertiary structures and are said to be intrinsically unstable. Previous research has demonstrated that phosphorylation sites

preferentially occur within regions of instability. Hence, the co-occurrence of phosphorylation sites could also be due to intrinsic instability.

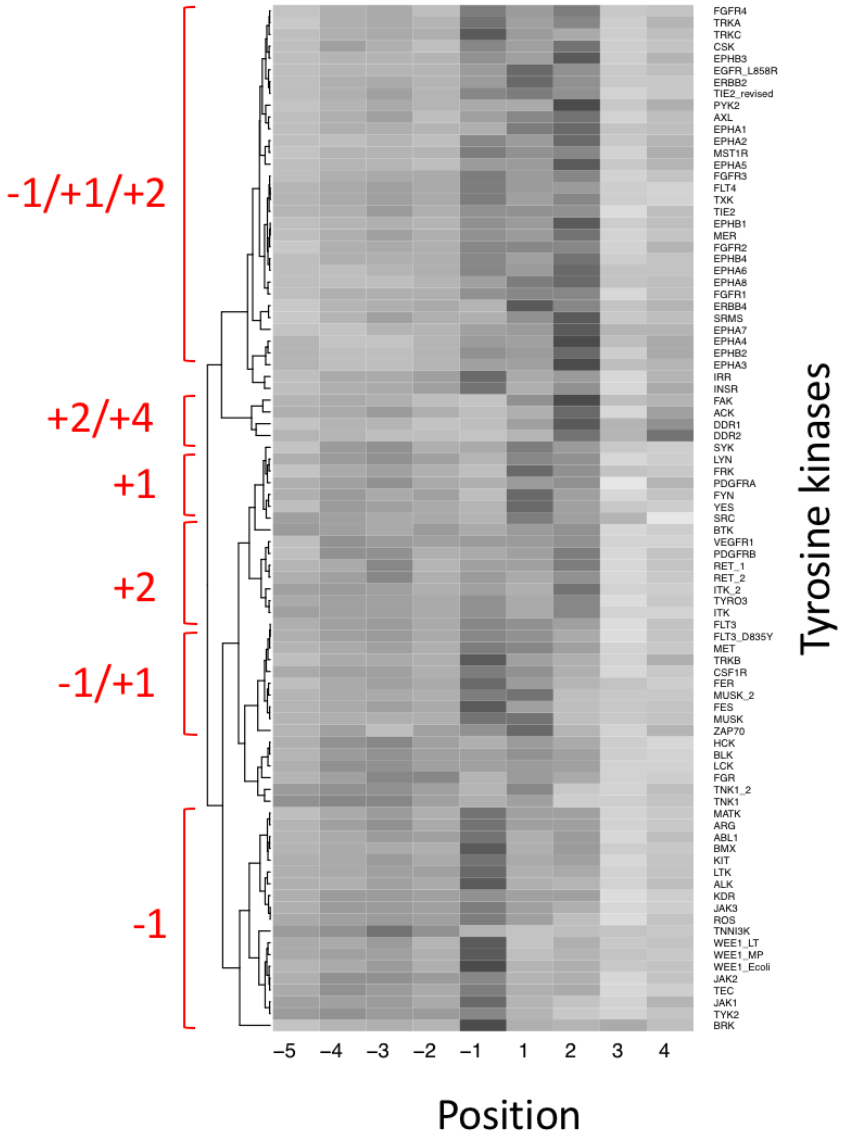

Figure 3-8: **Phospho-tyrosine selectivity.**
Phospho-tyrosine selectivity for each of the tyrosine kinases. Y-axis shows relative position to the central phosphorylated tyrosine and y-axis shows the kinase.

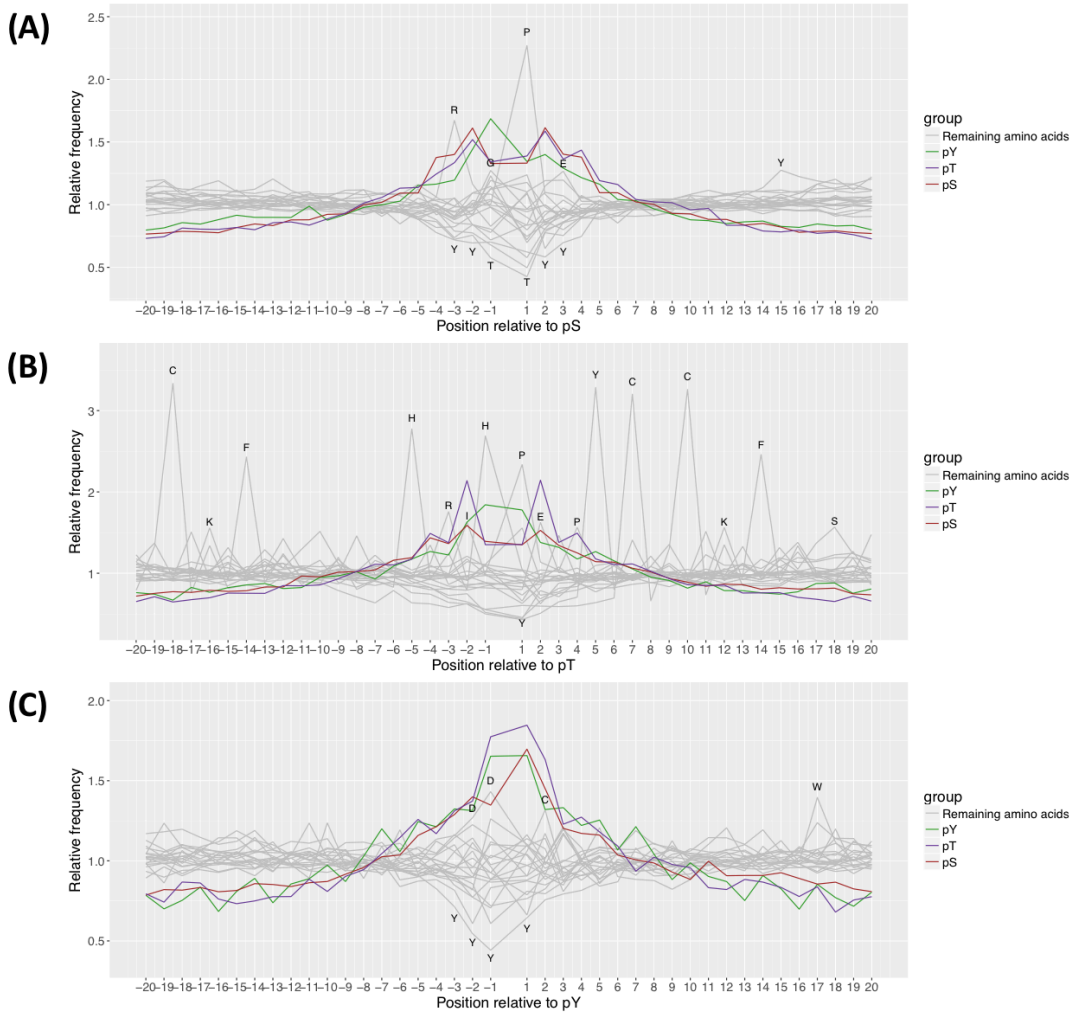

Figure 3-9: **Relative amino acid frequencies.**
(A) Phospho-serine, (B) phospho-threonine, and (C) phospho-tyrosine.

3.3 Summary

Presented in this chapter is the analysis of peptide library profiles of nearly 300 human protein kinases. I first demonstrated the quality of the data by showing the kinase specificity profiles cluster first by kinase type (tyrosine kinases and serine/threonine kinases) and then family. I also showed how the correlation between

kinase substrate specificity and the amino acid sequence of the kinase domain significantly increases when restricting to the DoS instead of full kinase domain. Both of these analyses support that the data is of sufficient quality for downstream analyses. I then used a simple empirical distribution for each kinase to benchmark the peptide library results to identify their reported substrates from PhosphoSite. I found that although the kinases certainly do better than a random model, there is room to improve the model when compared to other published algorithms that achieve higher accuracy. Finally, I explored phospho-priming to a small extent and found that kinases indeed have position-specific preferences for pY, pT, and pS. Moreover, using the PhosphoSite database, I computed the relative abundance of different amino acids (including pT, pY, and pS) around reported phosphorylation sites. I discovered a significant enrichment of phosphorylation sites near other phosphorylation sites. Hence, these two results indicate the potential significant relevance of phospho-priming in kinase signaling networks.

Determining the substrates of a kinase is a challenging problem, and one of the limiting factors behind understanding kinome-wide substrate specificity is having amino acid preferences on substrates for each kinase. Most kinases have few or no reported substrates, so building a model for their amino acid preferences is difficult to impossible. However, with the large peptide library dataset presented here that task is largely alleviated. But as my results demonstrate, kinase substrate specificity is also determined by other contextual cues like tissue/cellular expression and protein interaction networks in addition to amino acid content. Hence, further work will have to include such data to improve the predictions.

CHAPTER 4

CONCLUDING REMARKS

In this thesis I have demonstrated the utility of two orthogonal technologies for studying kinase function. Using next-generation sequencing, I found a large number of somatic alterations that affect protein kinases in the context of a GEM model of TNBC. I then focused on a number of those alterations and demonstrated their ability to drive tumorigenesis and vulnerability to therapeutic intervention. I then used the kinase peptide library from 262 protein kinases to discover their substrate specificity.

4.1 – Kinase drivers in triple-negative breast cancer

I used next-generation sequencing on a large set of primary tumors from a GEM model of TNBC. My colleague and I first validated that the GEM model reflects human TNBC. We confirmed using RNA-seq and IHC that most of the tumors are negative for HER2, PR, and ER, which are the primary markers of human TNBC. I also used the RNA-seq data to classify each tumor into one of the five intrinsic subtypes using two different algorithms. I found that the vast majority of tumors were of the basal subtype, which is the subtype most commonly associated with human TNBC. Finally, I demonstrated that the somatic alteration landscape of the tumors reflects human TNBC. I showed that the tumors have low mutations rates, but heterogenous CNA and gene fusion landscapes. Thus, we demonstrated that the GEM model is transcriptionally and genomically similar to human TNBC.

I then used a number of bioinformatics algorithms to discover the driver mutations, gene fusions, and CNAs. I found few recurrent alterations, except *Fgfr2* fusions and *Met*, *Yap1*, and *Myc* amplifications. A large number of the somatic alterations involved protein kinases. I found over a dozen mutations within protein kinase domains. Several protein kinases were found involved in in-frame gene fusions, including *Fgfr2*, *Braf*, *Raf1*, and *Met*. I also found CNAs involving protein kinases, including focally amplified *Met* in 10 tumors. In all, as has been shown in human breast cancer, protein kinases are a frequent source of tumor drivers.

We then selected a number of somatic alterations for follow-up validation, functional elucidation, and targeted therapy. We successfully validated for the functional relevance of the *Fgfr2*, *Braf*, and *Raf1* fusions as well as *Met*, *Pten*, *Egfr*, and *Fgfr2* CNAs. We selected drugs that specifically target each of the above gene fusions (including a Met overexpressing tumor) and were able to successfully either delay tumor growth or induce tumor regression for most tumors.

4.2 – Kinome substrate specificity

I used peptide library data for 262 kinases to predict the substrate specificity. I validated the quality of the data using unsupervised learning to show the peptide library data group the kinases according to kinase type and family. I then benchmarked the ability of the peptide library data to predict the kinases' putative substrates. I found that although the peptide library data performs better than random, there is room for improvement. Identifying or predicting a kinase's substrates using

either *in-vivo, in-vitro*, or *in-silico* methods is challenging. Benchmarking the peptide library data relies on a dataset of 'known' kinase-substrate pairs, such as from PhosphoSite or Phospho.ELM databases. Those databases scan the literature and even perform their own experiments (e.g. with mass-spec) for their data. However, it is entirely possible that a portion of those reported sites are false. Kinase signaling networks are intertwined and phosphorylation events can happen quickly, so in some cases a phosphorylation event could be assigned to one kinase while it was actually due to a downstream kinase. Furthermore, kinase substrate specificity depends on much more than amino acid composition. Cellular localization, tissue specificity, protein interaction networks, and more are also important, but were not incorporated into the model presented in this thesis.

4.3 – Future work

The presented analysis on mouse TNBC tumors could be furthered in several ways. Only a select few of the identified tumor drivers were chosen for follow-up validation and therapeutic targeting (*Fgfr2* fusions, *Braf* fusion, *Raf1* fusion, and *Met* over-expression). Many more driver alterations, especially those involving protein kinases, were discovered. Validation and targeting more of those alterations could lead to new ways of understanding TNBC tumorigenesis and discover new therapeutic strategies. TNBC is a very heterogenous disease with few recurrent alterations, so it is impractical to identify any drug that can be successful in a large number of tumors.

Hence, treatment needs to be tailored to the somatic alterations identified in individual tumors.

Determining which kinases phosphorylate which substrate is a complex problem. The algorithm presented in this work is a basic approach, and so there is significant room for improvement. The main route for improvement is the incorporation of contextual factors. Although amino acid composition is important for kinase specificity, it alone is insufficient for determining which kinase phosphorylates which substrate. Cellular localization, tissue specificity, protein interaction networks, and more are all important. Phospho-priming is also a ripe area for study using the peptide library data, and there are currently no computational analyses have taken a global look at phospho-priming (Valk et al., 2014).

Finally, there are a few areas of intersection between the two chapters presented in this work. Many of the somatic alterations identified in the mouse tumors involved protein kinases, including mutations, fusions, and CNAs. These alterations likely lead to changes in phosphorylation signaling networks, which can be measured using phospho-proteomics. The peptide library data could then be used to identify which kinases were responsible for the changing phosphorylation landscape.

BIBLIOGRAPHY

Ben-David,U. *et al.* (2016) The landscape of chromosomal aberrations in breast cancer mouse models reveals driver-specific routes to tumorigenesis. *Nat Commun*, **7**, 12160.

Beroukhim,R. *et al.* (2010) The landscape of somatic copy-number alteration across human cancers. *Nature*, **463**, 899–905.

Blake,J.A. *et al.* (2017) Mouse Genome Database (MGD)-2017: community knowledge resource for the laboratory mouse. *Nucleic Acids Research*, **45**, D723–D729.

Bórquez,D.A. and González-Billault,C. (2016) Bioinformatics Approaches for Predicting Kinase–Substrate Relationships. In, *Bioinformatics - Updated Features and Applications*. InTech, pp. 1–21.

Brandt,R. *et al.* (2000) Mammary gland specific hEGF receptor transgene expression induces neoplasia and inhibits differentiation. *Oncogene*, **19**, 2129–2137.

Brognard,J. and Hunter,T. (2011) Protein kinase signaling networks in cancer. *Current Opinion in Genetics & Development*, **21**, 4–11.

Campbell,J. *et al.* (2016) Large-Scale Profiling of Kinase Dependencies in Cancer Cell Lines. *Cell Rep*, **14**, 2490–2501.

Chakravarty,D. *et al.* (2017) OncoKB: A Precision Oncology Knowledge Base. *JCO Precision Oncology*, 1–16.

Chang,M.T. *et al.* (2016) Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. *Nat Biotechnol*, **34**, 155–163.

Chang,S.-S. *et al.* (2017) Aurora A kinase activates YAP signaling in triple-negative breast cancer. *Oncogene*, **36**, 1265–1275.

Chen,C. *et al.* (2014) Identification of a major determinant for serine-threonine kinase phosphoacceptor specificity. *Molecular Cell*, **53**, 140–147.

Cingolani,P. *et al.* (2014) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)*, **6**, 80–92.

Creixell,P., Palmeri,A., *et al.* (2015) Unmasking determinants of specificity in the human kinome. *Cell*, **163**, 187–201.

Creixell,P., Schoof,E.M., *et al.* (2015) Kinome-wide decoding of network-attacking mutations rewiring cancer signaling. *Cell*, **163**, 202–217.

Danecek,P. *et al.* (2012) High levels of RNA-editing site conservation amongst 15 laboratory mouse strains. *Genome Biol*, **13**, 26.

Davies,H. *et al.* (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, **534**, 47–54.

de Oliveira,P.S.L. *et al.* (2016) Revisiting protein kinase-substrate interactions: Toward therapeutic development. *Sci Signal*, **9**, re3–re3.

Dinkel,H. *et al.* (2011) Phospho.ELM: a database of phosphorylation sites--update 2011. *Nucleic Acids Research*, **39**, D261–7.

Easton,D.F. *et al.* (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**, 1087–1093.

Ferguson,F.M. and Gray,N.S. (2018) Kinase inhibitors: the road ahead. *Nat Rev Drug Discov*, **8**, 96.

Fleuren,E.D.G. *et al.* (2016) The kinome 'at large' in cancer. *Nat Rev Cancer*, **16**, 83–98.

Foulkes,W.D. *et al.* (2010) Triple-negative breast cancer. *N Engl J Med*, **363**, 1938–1948.

Frese,K.K. and Tuveson,D.A. (2007) Maximizing mouse cancer models. *Nat Rev Cancer*, **7**, 654–658.

Geyer,F.C. *et al.* (2017) Genetic analysis of microglandular adenosis and acinic cell carcinomas of the breast provides evidence for the existence of a low-grade triple-negative breast neoplasia family. *Mod. Pathol.*, **30**, 69–84.

Graveel,C.R. *et al.* (2009) Met induces diverse mammary carcinomas in mice and is associated with human basal breast cancer. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 12909–12914.

He,B. *et al.* (2009) Predicting intrinsic disorder in proteins: an overview. *Cell Res.*, **19**, 929–949.

Holstege,H. *et al.* (2010) Cross-species comparison of aCGH data from mouse and human BRCA1- and BRCA2-mutated breast cancers. *BMC Cancer*, **10**, 455.

Hornbeck,P.V. *et al.* (2004) PhosphoSite: A bioinformatics resource dedicated to physiological protein phosphorylation. *Proteomics*, **4**, 1551–1561.

Jang,J.S. *et al.* (2015) Common Oncogene Mutations and Novel SND1-BRAF Transcript Fusion in Lung Adenocarcinoma from Never Smokers. *Sci Rep*, **5**, 9755.

Katoh,M. (2016) FGFR inhibitors: Effects on cancer cells, tumor microenvironment and whole-body homeostasis (Review). *International Journal of Molecular Medicine*, **38**, 3–15.

Kinsella,R.J. *et al.* (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database*, **2011**, bar030–bar030.

Koboldt,D.C., Fulton,R.S., *et al.* (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.

Koboldt,D.C., Zhang,Q., *et al.* (2012) VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*, **22**, 568–576.

Kuilman,T. *et al.* (2015) CopywriteR: DNA copy number detection from off-target sequence data. *Genome Biol*, **16**, 49.

Lawrence,M.S. *et al.* (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, **505**, 495–501.

Lehmann,B.D. *et al.* (2011) Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.*, **121**, 2750–2767.

Lemmon,M.A. and Schlessinger,J. (2010) Cell Signaling by Receptor Tyrosine Kinases. *Cell*, **141**, 1117–1134.

Letai,A. (2017) Functional precision cancer medicine-moving beyond pure genomics. *Nature Medicine*, **23**, 1028–1035.

Liu,H. *et al.* (2017) Identifying and Targeting Sporadic Oncogenic Genetic Aberrations in Mouse Models of Triple-Negative Breast Cancer. *Cancer Discovery*, **8**, 354–369.

Liu,J.C. *et al.* (2014) Combined deletion of Pten and p53 in mammary epithelium accelerates triple-negative breast cancer with dependency on eEF2K. *EMBO Molecular Medicine*, **6**, 1542–1560.

Liu,X. *et al.* (2007) Somatic loss of BRCA1 and p53 in mice induces mammary tumors with features of human BRCA1-mutated basal-like breast cancer. *Proc Natl Acad Sci USA*, **104**, 12111–12116.

Maaten,L.V.D. and Hinton,G. (2008) Visualizing Data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605.

Manning,G. *et al.* (2002) The protein kinase complement of the human genome. *Science*, **298**, 1912–1934.

Marozkina,N.V. *et al.* (2008) MMTV-EGF receptor transgene promotes preneoplastic conversion of multiple steroid hormone-responsive tissues. *J. Cell. Biochem.*, **103**, 2010–2018.

Massard,C. *et al.* (2017) High-Throughput Genomics and Clinical Outcome in Hard-to-Treat Advanced Cancers: Results of the MOSCATO 01 Trial. *Cancer Discovery*, **7**, 586–595.

Masuda,H. *et al.* (2013) Differential Response to Neoadjuvant Chemotherapy Among 7 Triple-Negative Breast Cancer Molecular Subtypes. *Clinical Cancer Research*, **19**, 5533–5540.

Masuda,H. *et al.* (2012) Role of epidermal growth factor receptor in breast cancer. *Breast Cancer Res Treat*, **136**, 331–345.

Matissek,K.J. *et al.* (2018) Expressed Gene Fusions as Frequent Drivers of Poor Outcomes in Hormone Receptor-Positive Breast Cancer. *Cancer Discovery*, **8**, 336–353.

Newman,R.H. *et al.* (2014) Toward a systems-level view of dynamic phosphorylation networks. *Front Genet*, **5**, 263.

Nicorici,D. *et al.* (2014) **FusionCatcher**-a tool for finding somatic fusion genes in paired-end RNA-sequencing data. *biorxiv*.

Ornitz,D.M. *et al.* The fibroblast growth factor signaling pathway. *Wiley Online Library*.

Paquet,E.R. and Hallett,M.T. (2014) Absolute Assignment of Breast Cancer Intrinsic Molecular Subtype. *JNCI Journal of the National Cancer Institute*, **107**, dju357–dju357.

Parker,J.S. *et al.* (2009) Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology*, **27**, 1160–1167.

Pauli,C. *et al.* (2017) Personalized In Vitro and In Vivo Cancer Models to Guide Precision Medicine. *Cancer Discovery*, **7**, 462–477.

Pfefferle,A.D. *et al.* (2016) Genomic profiling of murine mammary tumors identifies potential personalized drug targets for p53-deficient mammary cancers. *Dis Model Mech*, **9**, 749–757.

Ponzo,M.G. *et al.* (2009) Met induces mammary tumors with diverse histologies and is associated with poor outcome and human basal breast cancer. *Proc Natl Acad Sci USA*, **106**, 12903–12908.

Reimand,J. *et al.* (2013) The mutational landscape of phosphorylation signaling in cancer. *Sci Rep*, **3**, 2651.

Rice,P. *et al.* (2000) EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.*, **16**, 276–277.

Sboner,A. and Elemento,O. (2016) A primer on precision medicine informatics. *Briefings in Bioinformatics*, **17**, 145–153.

Shah,S.P. *et al.* (2012) The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature*, **486**, 395–399.

Shaver,T.M. *et al.* (2016) Diverse, Biologically Relevant, and Targetable Gene Rearrangements in Triple-Negative Breast Cancer and Other Malignancies. *Cancer Res.*, 1–41.

Songyang,Z. *et al.* (1995) Catalytic specificity of protein-tyrosine kinases is critical for selective signalling. *Nature*, **373**, 536–539.

Songyang,Z. *et al.* (1994) Use of an oriented peptide library to determine the optimal substrates of protein kinases. *Curr. Biol.*, **4**, 973–982.

Stephens,P.J. *et al.* (2016) Complex landscapes of somatic rearrangement in human breast cancer genomes. *Nature*, **462**, 1005–1010.

Stransky,N. *et al.* (2014) The landscape of kinase fusions in cancer. *Nat Commun*, **5**, 4846.

Talevich,E. *et al.* (2016) CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol*, **12**, e1004873.

Taylor,S.S. and Kornev,A.P. (2011) Protein kinases: evolution of dynamic regulatory proteins. *Trends Biochem. Sci.*, **36**, 65–77.

Taylor,S.S. *et al.* (1995) How do protein kinases discriminate between serine/threonine and tyrosine? Structural insights from the insulin receptor protein-tyrosine kinase. *FASEB J.*, **9**, 1255–1266.

Ubersax,J.A. and Ferrell,J.E.,Jr. (2007) Mechanisms of specificity in protein phosphorylation. *Nature Reviews Molecular Cell Biology*, **8**, 530–541.

Valk,E. *et al.* (2014) Multistep phosphorylation systems: tunable components of biological signaling circuits. *Mol. Biol. Cell*, **25**, 3456–3460.

Wagih,O. *et al.* (2015) MIMP: predicting the impact of mutations on kinase-substrate phosphorylation. *Nat. Methods*, **12**, 531–533.

Wang,S. *et al.* (2016) Targeted Pten deletion plus p53-R270H mutation in mouse mammary epithelium induces aggressive claudin-low and basal-like breast cancer. *Breast Cancer Research*, **18**, 123.

Weinberg,R. (2013) The Biology of Cancer, Second Edition Garland Science.

Wu,Y.-M. *et al.* (2013) Identification of targetable FGFR gene fusions in diverse cancers. *Cancer Discovery*, **3**, 636–647.

Yoshihara,K. *et al.* (2015) The landscape and therapeutic relevance of cancer-associated transcript fusions. *Oncogene*, **34**, 4845–4854.

Yu,F.-X. *et al.* (2015) Hippo Pathway in Organ Size Control, Tissue Homeostasis, and Cancer. *Cell*, **163**, 811–828.

Zack,T.I. *et al.* (2013) Pan-cancer patterns of somatic copy number alteration. *Nature Publishing Group*, **45**, 1134–1140.

Zanconato,F. *et al.* (2016) YAP/TAZ as therapeutic targets in cancer. *Curr Opin Pharmacol*, **29**, 26–33.

Zhao,H. *et al.* (2014) CrossMap: a versatile tool for coordinate conversion between genome assemblies. *Bioinformatics*, **30**, 1006–1007.